МИНИСТЕРСТВО ЗДРАВООХРАНЕНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ
УЧРЕЖДЕНИЕ ОБРАЗОВАНИЯ «ВИТЕБСКИЙ ГОСУДАРСТВЕННЫЙ ОРДЕНА ДРУЖБЫ НАРОДОВ МЕДИЦИНСКИЙ УНИВЕРСИТЕТ»

С.Л. Гараничева, В.А. Таллер, Е.Г. Машеро

ОСНОВЫ СТАТИСТИКИ

Рекомендовано учебно-методическим объединением по высшему медицинскому, фармацевтическому образованию Республики Беларусь в качестве учебно-методического пособия для студентов учреждений высшего образования обучающихся по специальностям: 1-79 01 01 «Лечебное дело», 1-79 01 07 «Стоматология»

Витебск, 2019

УДК 31:004(072)
ББК 32.81я73+60.6я73
Г 20

Рецензенты:

М.Н. Борисевич – к. физ-мат. н., доцент, заведующий кафедрой компьютерного образования учреждения образования «Витебская ордена «Знак Почета» государственная академия ветеринарной медицины».

Кафедра прикладного и системного программирования учреждения образования «Витебский государственный университет имени П.М. Машерова» (кандидат физико-математических наук С.А. Ермоченко).

Гараничева, С. Л.
Г 20 **Основы статистики** : учеб.-метод. пособие / С. Л. Гараничева, В. А. Таллер, Е. Г. Машеро. – Витебск, ВГМУ, 2019. – 163 с.

ISBN 978-985-466-957-1

В учебно-методическом пособии изложены основные вопросы применения современных информационных компьютерных технологий для статистического анализа данных, представлен материал для выполнения студентами медицинского университета практических работ по основам статистики в среде электронных таблиц Microsoft Excel на втором году обучения. Пособие предназначено для студентов медицинских вузов, врачей-интернов, магистрантов, клинических ординаторов, аспирантов, слушателей курсов повышения квалификации и практических врачей.

УДК 31:004(072)
ББК 32.81я73+60.6я73

ISBN 978-985-466-957-1

© С.Л. Гараничева, В.А. Таллер, Е.Г. Машеро, 2019
© УО «Витебский государственный медицинский университет», 2019

ВВЕДЕНИЕ

В современном обществе к статистическим методам проявляется повышенный интерес как к одному из важнейших аналитических инструментариев в сфере поддержки процессов принятия решений. Большим шагом вперед к развитию статистической науки послужило применение экономико-математических методов и использование компьютерной техники в анализе медико-биологических процессов и социально-экономических явлений.

Стандартные статистические методы обработки данных включены в состав электронных таблиц, таких, как Lotus 1-2-3, QuattroPro, Microsoft Excel и др.; в математические пакеты общего назначения — Mathcad, Mathlab, Maple и т.д. Еще более мощными возможностями статистической обработки обладают специализированные русскоязычные пакеты — STADIA, МЕЗОЗАВР, СИГАМД, ОЛИМП:СтатЭксперт и др., и зарубежные – STATGRAPHICS, SPSS, SAS, BMDP, STATISTICA и др.

Наибольшее распространение в деловой сфере получил табличный процессор Microsoft Excel. За последние годы его популярность еще более возросла, что объясняется органичной интеграцией табличного процессора в пакет Microsoft Office.

Для проведения статистической обработки информации программа Microsoft Excel включает в себя надстройку пакет «Анализ данных» и библиотеку статистических функций. В повседневной деятельности такого набора инструментов, как правило, бывает вполне достаточно для проведения довольно полного и качественного статистического анализа информации. Если же пользователя не удовлетворяют подобные возможности Microsoft Excel, тогда необходимо обратиться к мощным специализированным пакетам статистического анализа, в частности к пакету STATISTICA фирмы StatSoft.

В данном учебно-методическом пособии будет использоваться следующее материальное оснащение:

Оборудование:

- IBM совместимые персональные компьютеры архитектуры X86.

Программное обеспечение и исходные данные (файлы):

1. Операционная система **Microsoft Windows**.
2. Программа **Microsoft Excel** из пакета **Microsoft Office**.
3. Файлы, соответствующие номеру выполняемой работы, с именами «Пр.зан.№k», где k – номер практической работы.

Методические материалы:

1. Гараничева, С.Л. Excel для студента-медика [Электронный ресурс] : учеб.-метод. пособие / С. Л. Гараничева. – Витебск : ВГМУ, 2012. – 1 электрон. опт. диск (CD ROM). – Excel для студента-медика. – ISBN 978-985-466-579-5. – 236 с.

ОГЛАВЛЕНИЕ

1. ПРИМЕНЕНИЕ СТАТИСТИЧЕСКОГО АНАЛИЗА ДАННЫХ В ХОДЕ НАУЧНЫХ ИССЛЕДОВАНИЙ И ДЛЯ РЕШЕНИЯ ПРОФЕССИОНАЛЬНЫХ ЗАДАЧ

Области применения статистического анализа данных в медицине и здравоохранении

В настоящее время в медицине и здравоохранении математические методы обработки данных широко применяются:

- для определения эффективности новых методик диагностики, лечения и реабилитации пациентов;
- при апробации новых фармакологических препаратов, эффективность действия которых на организм человека надо подтвердить или опровергнуть;
- при исследовании действия на измеряемую величину одного или нескольких факторов (например, степень влияния тяжести специального браслета на частоту самопроизвольного дрожания мышц рук) — дисперсионный анализ;
- для выявления степени взаимосвязи между отдельными явлениями и процессами, например между частотой сердечных сокращений (ударов в минуту) и частотой дыхания (вдохов в минуту) — корреляция;
- для выявления наиболее существенной периодической зависимости и их задержки в одном процессе или между несколькими процессами — корреляционный анализ;
- при нахождении периодических (и квазипериодических) зависимостей в данных — спектральный анализ (например, анализ ритмов в энцефалографии);
- для преобразования временных рядов с целью удаления из них высокочастотных и низкочастотных колебаний — сглаживание и фильтрация (например, фильтрация электрокардиограмм с целью удаления артефактов и помех);
- в других направлениях обработки медико-биологических данных.

Следует отметить, что материал данного пособия ориентирован на первичное ознакомление студентов с практическими приемами проведения элементарной статистической обработки данных. Теоретические знания по данной теме студенты получают на кафедре медицинской и биологической физики при изучении дисциплины «Медицинская и биологическая физика». Более глубокие знания по данной теме студенты лечебного факультета смогут получить при изучении дисциплины «Общественное здоровье и здраво-

охранение», которая преподается на одноименной кафедре на 4-6 курсах медицинского вуза, в темах «Медицинская статистика» и «Методы статистической обработки данных».

В основе обработки и анализа данных лежат математические методы, которые в большинстве своем являются неизменными в течение десятилетий. Компьютерный анализ медико-биологических данных предполагает обработку данных с помощью определенных программных средств. Следовательно, специалисту-медику необходимо иметь представление как о самих математических методах анализа данных, так и о программных средствах, реализующих эти методы.

В связи с тем, что в данном курсе мы на углубленном уровне не изучаем приемы работы с приложениями интегрированного пакета Microsoft Office, напомним возможности электронных таблиц (ЭТ) Microsoft Excel.

ЭТ Microsoft Excel позволяют выполнить:

- автоматизацию вычислений с помощью подготовленных макетов;
- построение диаграмм и графиков;
- обработку данных списков — простейших баз данных;
- создание и исследование медико-биологических моделей;
- решение задач прогнозирования;
- статистическую обработку медико-биологических данных;
- аппроксимацию графиков зависимостей данных полученных эмпирическим путем математическими функциями (т.е. получение математических формул, для описания функциональных зависимостей);
- решение задач оптимизации и др.

Для лучшего восприятия материала вначале ознакомимся с основными сведениями по статистике, а затем рассмотрим возможности их реализации в среде электронных таблиц.

Вопросы для самоконтроля

1. Укажите области применения статистического анализа данных в медицине и здравоохранении.
2. Приведите примеры применения статистического анализа.
3. В каких дисциплинах, изучаемых Вами ранее в медицинском вузе, целесообразно использовать статистический анализ?
4. Какие изучаемые Вами ранее программы позволяют выполнять статистический анализ данных?

2. ОСНОВНЫЕ ПОНЯТИЯ МЕДИЦИНСКОЙ СТАТИСТИКИ

2.1. ТЕОРЕТИЧЕСКИЕ ОСНОВЫ

Случайное событие — событие, которое может произойти или не произойти без видимой закономерности.

Случайная величина — величина, принимающая различные значения без видимой закономерности, т.е. случайным образом.

Вероятность (p) — параметр, характеризующий частоту появления случайного события. Вероятность изменяется от 0 до 1, причем вероятность p = 0 означает, что случайное событие никогда не произойдет, вероятность p = 1 означает, что случайное событие происходит всегда.

Переменная — любая варьируемая величина (исследуемый признак).

Генеральная совокупность — множество единиц наблюдения, охватываемых сплошным наблюдением.

В связи с тем, что невозможно провести анализ необходимых признаков во всей генеральной совокупности, исследование генеральной совокупности заменяют исследованием **выборки**. Выборочный метод используется для получения правильных выводов относительно всей совокупности объектов. Конечной целью изучения выборки всегда является получение информации о генеральной совокупности**.**

Выборка — группа элементов, выбранная для исследования из всей совокупности элементов. Задача выборочного метода состоит в том, чтобы сделать правильные выводы относительно всего собрания объектов, их совокупности. Например, врач делает заключение о составе крови пациента на основе анализа ее нескольких капель. Выборка может включать один или несколько признаков, которые в информатике называются переменными.

При формировании выборки должен выполняться ряд условий.

1. Каждый член генеральной совокупности должен иметь равную вероятность попасть в выборку.
2. Отбор единиц из генеральной совокупности необходимо производить независимо от изучаемого признака.
3. Отбор должен проводиться из однородных групп (например, одинаковое соотношение полов).

Случайная выборка формируется случайным образом — наудачу.

Чтобы по выборке можно было судить о генеральной совокупности, выборка должна быть представительной (репрезентативной). Репрезентативность означает, что объекты выборки достаточно хорошо представляют генеральную совокупность. Различают *количественную* и *качественную* репрезентативность.

Количественная репрезентативность определяется *числом наблюдений,* гарантирующих получение статистически достоверных данных. Чем больше наблюдений, тем результаты достоверней.

Качественная репрезентативность обозначает *структурное соответствие* выборочной и генеральной совокупностей, например, по половозрастным признакам.

Виды частотных распределений. Свойства частотного распределения Гаусса

Выборка описывается рядом параметров, среди которых: закон частотного распределения элементов в выборке, среднее арифметическое, медиана, мода, дисперсия, стандартное отклонение, максимальное и минимальное значения, асимметрия, эксцесс и другие величины.

Элементы выборки обычно распределены в соответствии с каким-то законом, который чаще всего можно описать математически (биномиальное, распределение Пуассона, нормальное распределение, распределение Фишера и др.). Для исследователя является принципиальным, подчиняются ли элементы выборки нормальному закону частотного распределения или какому-то другому.

Нормальное распределение (распределение Гаусса) используется для приближенного описания явлений, в которых на результат воздействует большое количество независимых случайных факторов, имеющих вероятностный **(**случайный) характер, среди которых нет сильно выделяющихся. В связи с тем, что подавляющее большинство медико-биологических явлений носит вероятностный характер, нормальное распределение в этих явлениях встречается весьма часто и имеет вид, представленный на рисунке 1.

Таким образом, ***нормальное частотное распределение*** описывает совокупность объектов, в которой крайние значения некоторого признака — наименьшее и наибольшее — появляются редко; чем ближе значение признака к среднему арифметическому, тем чаще оно встречается. Например, распределение пациентов по их чувствительности к воздействию любого фармакологического агента часто приближается к нормальному распределению.

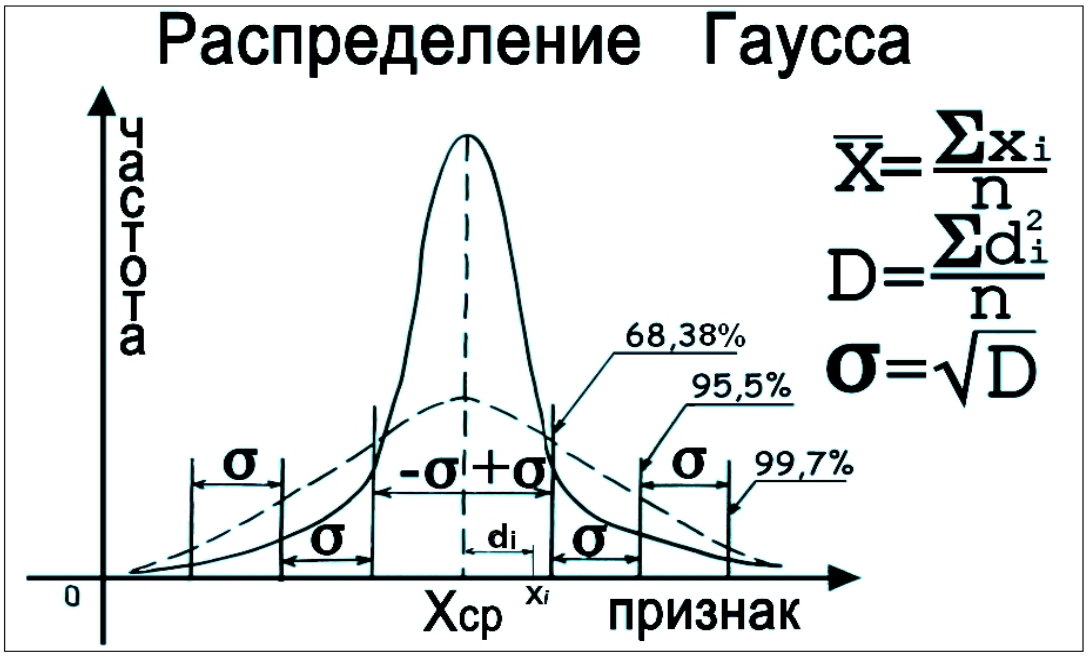


Рис. 1. Вид распределения Гаусса

При большом количестве объектов в выборке (n>300) любой закон частотного распределения, описывающий эту выборку, стремится к нормальному.

Гистограмма выборки — графическое изображение зависимости частоты попадания элементов выборки в заданные интервалы исследуемого признака от соответствующего интервала группировки. Например, можно построить графическое изображение числа случаев смерти от инфаркта миокарда в зависимости от возраста умерших.

Описательная статистика

Совокупность статических характеристик, описывающих выборку, называют ***описательной статистикой***. Рассмотрим основные статистические характеристики выборки.

Среднее значение (M) — центр выборки, вокруг которого группируются элементы выборки. Среднее значение в литературе обозначается $x_{ср}$, $\bar{x}$, A.

✓ **Обратите внимание!** В нашем случае символом **M** в параметрах функций обозначен массив данных, т.е. диапазон ячеек в которых он расположен.

Медиана (M) — элемент выборки, число элементов выборки со значениями больше которого и меньше которого — равно.

Мода (M) — элемент выборки с наиболее часто встречающимся значением.

Дисперсия (M) (обозначается ***D)*** — параметр, характеризующий степень разброса элементов выборки относительно среднего значения. Чем больше дисперсия, тем дальше отклоняются значения элементов выборки от среднего значения.

Стандартное отклонение (M) — среднее квадратическое отклонение, обозначается ***S*** — параметр аналогичный дисперсии, но имеющий ту же размерность, что и среднее значение, а поэтому и более удобный.

Ошибка среднего или стандартная ошибка (m или $\Delta x_{ср}$***)*** — параметр, характеризующий степень возможного отклонения среднего значения, полученного на исследуемой ограниченной выборке, от истинного среднего значения, полученного на всей совокупности элементов.

Асимметрия (M) — величина, характеризующая несимметричность распределения элементов выборки относительно среднего значения. Коэффициент асимметрии в случае симметричного распределения равен 0. Если его значение отрицательно, то частотное распределение имеет более длинный левый хвост чем нормальное, если положительно — то распределение имеет более длинный правый хвост чем нормальное.

Эксцесс (M) — характеризует островершинность пика распределения. Коэффициент эксцесса в случае остроты вершины, соответствующей нормальному распределению, равен 0. Если эксцесс положителен, то пик

заострен, если отрицателен — соответственно, закруглен.

Часто значения асимметрии и эксцесса используют для проверки гипотезы о том, что наблюдаемые данные (выборка) соответствуют нормальному распределению. Если коэффициенты асимметрии и эксцесса по модулю не превышают значение ***0,1 (|Ка| ≤ 0,1 и |Кэ| ≤ 0,1)*** частотное распределение близко к нормальному.

Среднее арифметическое **(** $\bar{x}$ **)** возвращает функция Microsoft Excel СРЗНАЧ(М), где М массив выборки. Это центр выборки, вокруг которого группируются элементы выборки (рис. 1). Оно выражает характерную, типичную для данного ряда величину признака и дает возможность:

1) точечно оценить генеральный параметр математического ожидания случайной величины, т.е. охарактеризовать исследуемую совокупность одним числом;

2) сравнить отдельные величины со средним арифметическим;

3) определить тенденцию развития какого-либо явления;

4) сравнить разные совокупности;

5) вычислить другие статистические показатели, так как многие статистические вычисления опираются на среднее арифметическое.

Недостатки среднего арифметического

Так как среднее арифметическое чувствительно к воздействию экстремальных (очень больших или очень малых) значений, оно неприменимо для описания асимметрично распределенных данных (смещенных распределений). Для этого лучше подходит медиана.

✓ **Обратите внимание!** Среднее арифметическое целесообразно вычислять только для количественных признаков.

Дисперсия (D), характеризует разброс параметров выборки вокруг среднего значения. Значение дисперсии вычисляется по формуле:

$$D = \sum_{i=1}^{n} d_i^2 / n .$$

Для увеличения точности формулу изменяют следующим образом

$$D = \sum_{i=1}^{n} d_i^2 /(n-1) ,$$

где n — число наблюдений,

d_i — отклонение варианты выборки от среднего.

Существенным недостатком дисперсии, которая является именованной величиной, является несоответствие ее размерности и размерности отдельных вариант числового ряда. Размерность дисперсии равна квадрату размерности исследуемого параметра.

Указанного недостатка лишено ***стандартное отклонение*** (σ, S — для выборки**)**. σ = КОРЕНЬ(D). Его значение вычисляет функция Microsoft Excel СТАНДОТКЛОН(М), где М — массив выборки. Геометрическая ин-

терпретация этого параметра представлена на рисунке 1.

В области, где располагается часть вариант (объектов выборки), отклоняющихся от среднего не более чем на **σ,** при нормальном распределении всегда оказывается 68,38 % всех вариант. В пределах от **-2σ** до **+2σ** лежат 95,5% всех вариант. В пределах от **-3σ** до **+3σ** находится 99, 7% всех вариант выборки (правило трех сигм).

На рисунке 2 представлен пример вычисления средствами приложения Microsoft Excel среднего значения и стандартного отклонения.

При вычислении среднего значения выборки, стандартного отклонения (рис. 2) в качестве массива в соответствующих функциях Microsoft Excel используются массивы (интервалы размещения значений элементов выборок): *B4:B10* — для контрольной группы и *C4:C10* — для исследуемой группы.

B11 =СРЗНАЧ(B4:B10)

	A	B	C	D
1	***Результаты исследования частоты сердечных сокращений (ЧСС)***			
2		***Группы***		
3	***Статистические характеристики***	***Контрольная***	***Исследуемая***	
4		162	135	
5		156	126	
6		144	115	
7		137	140	
8		125	121	
9		145	112	
10		151	130	
11	***Среднее значение***	***145,7142857***		
12	***Стандартное отклонение***	***12,29788987***		
13				

Рис. 2. Вид Рабочего листа ЭТ Microsoft Excel с примерами вычисления среднего значения и стандартного отклонения выборки

Функции, родственные функции МЕДИАНА (М)

Функция КВАРТИЛЬ(М;k), где *М* массив, *k* – часть.

Синтаксис: ***КВАРТИЛЬ(массив; часть)***

Результат: *Рассчитывает квартиль для множества данных.*

Аргументы:

– *массив*: диапазон ячеек с числовыми значениями, для которых определяются значения квартилей;

– *часть*: аргумент, определяющий, что будет рассчитывать функция КВАРТИЛЬ().

Примечание: функция КВАРТИЛЬ() рассчитывает:

- минимальное значение, если аргумент *часть* = 0;
- первый квартиль (25 %), если аргумент *часть* = 1;

- значение медианы (50 %), если аргумент *часть* = 2;
- третий квартиль (75 %), если аргумент *часть* = 3;
- максимальное значение, если аргумент *часть* = 4;

Функции *МИН(), МЕДИАНА() и МАКС()* рассчитывают то же самое значение, что и функция *КВАРТИЛЬ()*, если аргумент *часть* равен 0, 2 или 4 соответственно.

Квартили представляют собой значения признака, делящие ранжированную совокупность на четыре равновеликие части. Различают квартиль нижний *(Q_1),* отделяющий 1/4 часть совокупности с наименьшими значениями признака, и квартиль верхний (Q_3), отделяющий 1/4 часть с наибольшими значениями признака. Средним квартилем (*Q_2)* является медиана.

Функция КВАРТИЛЬ() не требует предварительной ранжировки данных, она проводит ее автоматически.

Функция ПЕРСЕНТИЛЬ(M,k)

Кроме квартилей в вариационных рядах распределения могут определяться децили и персентили. Последние также иногда называют перцентилями или процентилями. ***Децили*** делят ранжированную совокупность на ***десять*** равновеликих частей, а ***персентили*** — на ***сто***. Персентили применяются лишь при необходимости подробного изучения структуры вариационного ряда.

ПЕРСЕНТИЛЬ(массив; k)

Результат: *Рассчитывает* ***k****-й персентиль для множества данных.*

Аргументы:

- *массив*: диапазон ячеек с числовыми значениями, для которых определяются значения персентилей;
- *k:* значение персентиля в интервале от 0 до 1 включительно.

Замечание: *если массив пуст или содержит более 8191 точки данных, то функция ПЕРСЕНТИЛЬ() помещает в ячейку значение ошибки #ЧИСЛО!*

Понятие доверительного интервала

Интервал, в котором с заданной доверительной вероятностью ***P=1 – α***, где α — уровень значимости, находится оцениваемый параметр, называется ***доверительным интервалом*.**

Как отмечалось выше, стандартная ошибка среднего (обозначается буквами **m** или $\mathbf{\Delta x_{ср}}$**)** характеризует интервал значений [$\mathbf{x_{ср} - \Delta x_{ср}}$ **;** $\mathbf{x_{ср} + \Delta x_{ср}}$], в котором предположительно находится среднее генеральной совокупности.

При работе с выборками небольшого объема, точечная оценка (среднее, медиана, мода) в генеральной совокупности и выборке могут существенно различаться. В этом случае вместо точечных оценок обычно пользуются интервальными.

Интервальные оценки задают двумя выборочными значениями – концами интервала, в котором для генеральной совокупности оказывается параметр (среднее, медиана, мода) с некоторой заранее оговоренной вероятностью**.**

Обычно при определении доверительного интервала используют доверительную вероятность равную 95%. Половина доверительного интервала называется предельной ошибкой среднего, обозначается $\mathbf{\Delta x_{пр}}$ и будет равна $\mathbf{\Delta x_{пр} = Kt * \Delta x_{ср}}$**,** где **Kt** — коэффициент, величина которого зависит от доверительной вероятности и объема выборки. Значение этого коэффициента вычисляется с помощью встроенной функции СТЬЮДРАСПОРБР(). На больших выборка при доверительной вероятности 95% это значение равно 1,96.

Итак, доверительный интервал среднего **–** это интервал значений исследуемого признака, $[X_{ср} - \Delta x_{пр}; X_{ср} + \Delta x_{пр}]$, на котором с доверительной вероятностью (обычно 95%) в генеральной совокупности находится среднее значение.

Формат представления рассчитанного доверительного интервала следующий:

M (95% ДИ: j…q) или Me (95% ДИ: j…q), например: M = 96,6 (95% ДИ: 90,2…101,4).

Вместо ***троеточия*** можно использовать ***дефис***, как в (95% ДИ: 90,2–101,4) – данный формат еще не устоялся, у него нет стандарта de facto. Тем не менее, в западной научной литературе указание доверительного интервала вместо стандартного отклонения – общепринятая практика; там ДИ обозначают как CI, т.е. «Confidence Interval».

Контрольная и исследуемая группы

При проведении научных исследований обычно используют несколько выборок (групп).

Контрольная группа — это выборка элементов генеральной совокупности, которая отражает исходные свойства этой совокупности.

Исследуемая группа — это совокупность объектов, над которыми проводится исследование. Она отражает свойства генеральной совокупности после воздействия на нее каким-либо конкретным фактором.

Основные критерии нормальности закона частотного распределения выборки

Целью изучения описательной статистики выборки является не только определение основных числовых значений параметров выборки, но и выявление закона ее частотного распределения. Если закон частотного распределения является нормальным, т.е. соответствует распределению Гаусса, то в дальнейшем для статистического анализа данных будут применяться параметрические методы обработки, в противном случае — непараметрические.

Параметрические методы являются более точными, поэтому частотное распределение стремятся преобразовать таким образом, чтобы оно соответствовало нормальному закону. С этой целью значение признака выборки логарифмируют или потенцирует, возводят в квадрат или извлекают корень квадратный, т.е. выполняют такое преобразование, которое позволяет приблизить закон частотного распределения к нормальному. Если соответствующее преобразование подобрать не удается, то используют менее точные непараметрические методы обработки данных.

Для того, чтобы по значению основных параметров выборки можно было судить о нормальности частотного распределения, разработан ряд приближенных практических критериев, среди них следующие.

1**.** При нормальном частотном распределении **среднее значение, медиана, мода** выборки приблизительно **равны**.

$$X_{ср} \sim Медиана \sim Мода.$$

2. Если **коэффициенты эксцесса — Кэ и ассиметрии — Ка и их стандартные ошибки** имеют один и **тот же порядок**, то частотное распределение можно считать нормальным.

3. В соответствии **с законом Колмогорова-Смирнова**, анализирующем частотное распределение:

- на расстоянии S от $X_{ср}$ находится ~68,38% элементов выборки;
- на расстоянии 2S от $X_{ср}$ находится ~ 95,5% элементов выборки;
- на расстоянии 3S от $X_{ср}$ находится ~ 99,7% элементов выборки.

4. В соответствии **с законом больших чисел,** если количество объектов **n** в выборке **очень большое** ($n \geq 300$) и исследуемый признак является количественной величиной, то частотное распределение выборки соответствует нормальному.

Краткие выводы

Для того, чтобы правильно использовать встроенные функции, следует с ними ознакомиться, воспользовавшись встроенной справочной системой **Microsoft Excel** раздела ***Справка ⇒ Обзор справки Excel ⇒ Справочник по функциям ⇒ Статистические функции***.

Наиболее важной информацией, позволяющей правильно применять функции, является информация: о **назначении** функции, **синтаксис** функции (перечень ее параметров) и **пример** использования функции.

Следует отметить, что элементарную статистическую обработку медико-биологических данных можно также выполнить с помощью инструментов надстройки Microsoft Excel — **пакет Анализ данных**, в котором представлены процедуры формирования и анализа свойств выборки и основные параметрические методы статистической обработки данных.

2.2. ПРИМЕНЕНИЕ СТАТИСТИЧЕСКИХ ФУНКЦИЙ MICROSOFT EXCEL ДЛЯ ВЫЧИСЛЕНИЯ ОСНОВНЫХ СТАТИСТИЧЕСКИХ ХАРАКТЕРИСТИК ВЫБОРКИ. ЧАСТОТНЫЙ АНАЛИЗ ДАННЫХ

ЦЕЛИ ЗАНЯТИЯ

1. Ознакомиться с основными характеристиками выборки.
2. Научиться вычислять описательные статистики выборки с помощью встроенных функций Microsoft Excel и процедуры «Описательная статистика» пакета Анализ данных.
3. По полученным результатам научиться формулировать выводы о нормальности частотного распределения выборки.
4. Научиться выбирать и обосновывать методы дальнейшей статистической обработки данных: параметрические, непараметрические.

МЕТОДИКА ВЫПОЛНЕНИЯ РАБОТЫ

Постановка задачи

Дано: Выборочная совокупность, содержащая сведения о частоте сердечных сокращений пациентов двух групп: контрольной и исследуемой.

Контрольная	***Исследуемая***
162	135
156	126
120	100
130	125
135	113
144	115
137	140
125	121
145	112
151	130

Требуется:

1. Изучить основные характеристики выборки, реализуемые встроенными функциями Excel, используя справочную систему программы.
2. Записать в таблицу 1 назначение, синтаксис каждой функции, кратко пример ее применения.
3. Вычислить основные характеристики двух выборок, с помощью встроенных функций.
4. Используя один из известных критериев нормальности частотного распределения, сделать вывод о нормальности частотного распределения в группах.
5. Обосновать выбор методов дальнейшей статистической обработки данных (параметрический, непараметрический).
6. Построить с помощью Мастера диаграмм программы Microsoft Excel гистограммы частотного распределения выборок.

ХОД РАБОТЫ

Вычисление описательной статистики выборки средствами встроенных функций Microsoft Excel

1. Скопируйте в свою папку из папки ***Z:\Материалы для работы\Статистика*** книгу Microsoft Excel с именем *Пр.зан.№1–Описательные Excel.xls.*

2. Переименуйте скопированный файл, задав в качестве имени файла номер практической работы и свою фамилию, номер группы. Например,

Пр.зан.№1 Иванов А. — 24 леч.

3. Изучите функции, приведенные в таблице 1, используя справочную систему программы, команда: ***Справка ⇒ Обзор справки Excel ⇒ Справочник по функциям ⇒ Статистические функции.***

4. Запишите для каждой функции, найденные в справке данные, в тетради в таблицу 1.

Таблица 1.

Характеристика основных встроенных функций Microsoft Excel для получения описательной статистики выборки

Название функции	Назначение функции	Синтаксис	Пример
СРЗНАЧ()	Вычисляет среднее значение по выборке	**СРЗНАЧ(A1:A10;B2)**	**СРЗНАЧ(A1:A10;B2;F3)**
*** m =S/Корень(n)**		-	
МЕДИАНА()			
МОДА()			
СТАНДОТКЛОН()			
ДИСП()			
ЭКСЦЕСС()			
СКОС()			
*** МАКС()-МИН()**			
МИН()			
МАКС()			
СУММ()			
СЧЕТ()			
ДОВЕРИТ()			
Уровень надежности (95%) ($\Delta x_{пр}$)			

5. Вычислите основные характеристики выборок, применив при вводе функций Мастер функций . Результаты запишите в тетради в таблицу 2.

✓ **Обратите внимание!** Значения в строках, помеченных звездочкой, следует вычислять по данным таблицы после ее заполнения. Формулы для вычисления приведены в таблице на Рабочем листе Microsoft Excel внизу (с. 141).

Таблица 2.

Описательная статистика выборок

Название функции	Контрольная группа	Исследуемая группа
СРЗНАЧ()		
*** m=S/Корень(n)**		
МЕДИАНА()		
МОДА()		
СТАНДОТКЛОН()		
ДИСП()		
ЭКСЦЕСС()		
СКОС()		
*** МАКС()-МИН()**		
МИН()		
МАКС()		
СУММ()		
СЧЕТ()		
ДОВЕРИТ()		
КВАРТИЛЬ (Массив;1)		
КВАРТИЛЬ (Массив;3)		
ПЕРСЕНТИЛЬ (Массив;0,15)		
ПЕРСЕНТИЛЬ (Массив;0,85)		
(СТЬЮДРАСПОБР)		
Уровень надежности (95%) ($\Delta x_{пр}$)		
Нижняя граница 95% ДИ для среднего		
Верхняя граница 95% ДИ для среднего		

6. Используя критерии нормальности, сделайте и обоснуйте выводы о нормальности частотных распределений в каждой из исследуемых групп. Выводы запишите в тетрадь в отчет по практической работе.

7. Обоснуйте, какие методы следует применять для дальнейшей обработки данных (параметрические или непараметрические). Запишите обоснование в тетрадь.

Построение гистограмм частотных распределений выборок

1. Скопируйте исходные данные Рабочего листа *«Задание 1»* на лист *«Задание 2»* книги Microsoft Excel, измените текст заголовка (ячейка B1) в соответствии с рисунком 3, удалите из ячеек лишнюю текстовую информацию.

2. Упорядочите (отсортируйте исходные данные) по возрастанию значений в каждой выборке. Отдельно выделяя диапазон каждой выборки, примените команду: ***Данные ⇒ Сортировка ⇒ По возрастанию ⇒ Сортировать в пределах указанного выделения.*** У вас должна получиться таб-

лица, представленная на рисунке 4.

	A	B	C
1		Построение гистограммы	
2		Группы	
3		Контрольная	Исследуемая
4		162	135
5		156	126
6		120	100
7		130	125
8		135	113
9		144	115
10		137	140
11		125	121
12		145	112
13		151	130
14			

Рис. 3. Данные после копирования

	A	B	C
1		Построение гистограммы	
2		Группы	
3		Контрольная	Исследуемая
4		120	100
5		125	112
6		130	113
7		135	115
8		137	121
9		144	125
10		145	126
11		151	130
12		156	135
13		162	140
14			

Рис. 4. Данные после сортировки

3. В ячейках E3:E10 приведены диапазоны значений признака, по которым следует вычислить частоты встречаемости исследуемого признака (ЧСС) (рис. 5).

4. Заполните частоты встречаемости ЧСС в этих диапазонах для контрольной и исследуемой группы.

	A	B	C	D	E	F	G
1		Построение гистограммы					
2		Группы					
3		Контрольная	Исследуемая		Диапазоны значений	Частота	
4		120	100			Контрольная	Исследуемая
5		125	112		100-119	0	4
6		130	113		120 - 129	2	3
7		135	115		130 - 139	3	2
8		137	121		140 - 149	2	1
9		144	125		150 - 159	2	0
10		145	126		160 -169	1	0
11		151	130				
12		156	135				
13		162	140				
14							

Рис. 5. Вид Рабочего листа Excel, подготовленного для заполнения частот по выбранным диапазонам

5. На основании полученных частот встречаемости признака постройте гистограммы частотных распределений для двух групп с помощью Мастера диаграмм Microsoft Excel. Примерный вид полученных гистограмм представлен на рисунке 6.

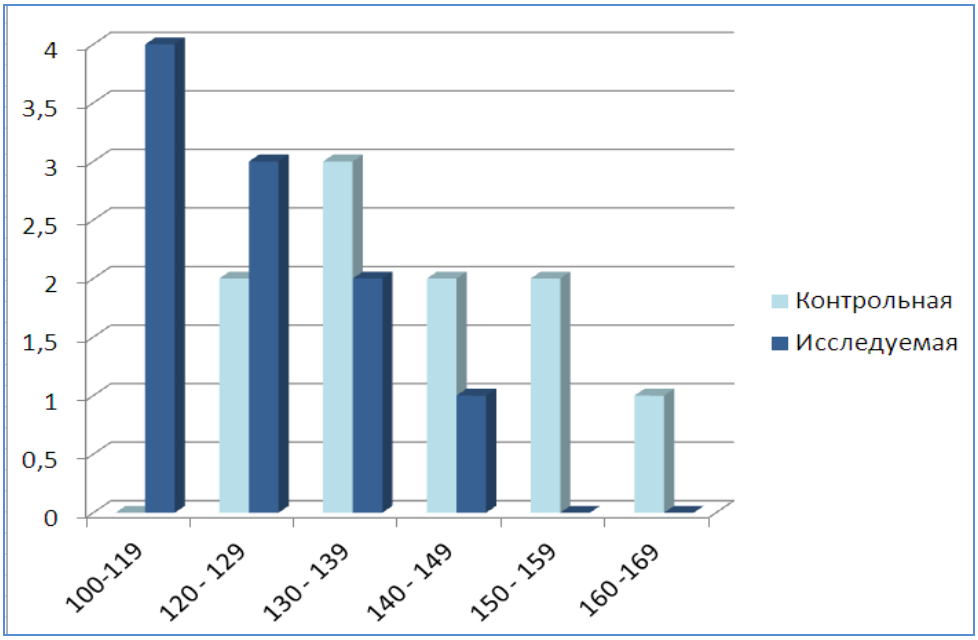


Рис. 6. Примерный вид гистограмм частотных распределений

6. По виду полученных гистограмм сделайте и запишите в тетрадь выводы о нормальности частотных распределений в двух группах.

7. Запишите вывод: в какую сторону сместились частоты в исследуемой группе по сравнению с контрольной, наблюдается ли улучшение процесса лечения?

ДОПОЛНИТЕЛЬНОЕ ЗАДАНИЕ

1. Используя полученные значения коэффициентов асимметрии и эксцесса, для каждой выборки проанализируйте и запишите в тетрадь для контрольной и исследуемой группы:

- в какую сторону смещено частотное распределение выборки (влево, вправо) относительно нормального?
- является ли вершина частотного распределения выборки более острой или пологой по сравнению со стандартным частотным распределением?

2. Сравните Ваши выводы с видом частотных распределений на гистограмме, соответствуют ли они?

Вопросы для самоконтроля

1. Что понимают под генеральной совокупностью?
2. Для каких целей формируется выборка?
3. Как следует формировать выборку из генеральной совокупности?
4. Что означает понятие репрезентативная выборка?

5. Какие характеристики позволяют описать выборку?

6. В чем различие между средним значением выборки и ее медианой, модой?

7. Для каких целей можно применять среднее арифметическое значение?

8. Когда для описания центральной тенденции выборки следует применять среднее, медиану?

9. Какие характеристики выборки получают при вычислении описательных статистик?

10. Какой показатель описательной статистики характеризует симметричность частотного распределения выборки?

11. Какой показатель описательной статистики характеризует островершинность частотного распределения выборки?

12. Какие процессы обычно описывает нормальное частотное распределение?

13. Назовите свойства нормального частотного распределения выборки.

14. Перечислите основные критерии нормальности частотного распределения выборки.

15. Для каких целей определяют нормальность частотного распределения выборки?

16. Какие методы статистического анализа (параметрические или непараметрические) являются наиболее точными?

17. Приведите формулы для вычисления стандартной ошибки среднего и предельной ошибки среднего.

18. В чем различие понятий стандартная ошибка среднего и предельная ошибка среднего?

19. Объясните сущность понятия доверительного интервала среднего, рассчитанного с вероятностью 95%.

20. Как вычисляют границы доверительного интервала для среднего (нижнюю, верхнюю)?

21. Что понимают под уровнем надежности?

3. ВОЗМОЖНОСТИ И ОСНОВНЫЕ ПРОЦЕДУРЫ ПАКЕТА АНАЛИЗ ДАННЫХ — НАДСТРОЙКИ MICROSOFT EXCEL

3.1. ТЕОРЕТИЧЕСКИЕ ОСНОВЫ СТРУКТУРЫ ПАКЕТА АНАЛИЗ ДАННЫХ

Частотный анализ данных можно выполнять как вручную, так и с помощью процедуры «Гистограмма». Эта процедура наряду с другими представлена в надстройке Microsoft Excel Анализ данных.

Технология работы в надстройке Анализ данных

Выберите в меню ***Данные*** пункт **Анализ данных** (в правом верхнем углу) (рис. 7).

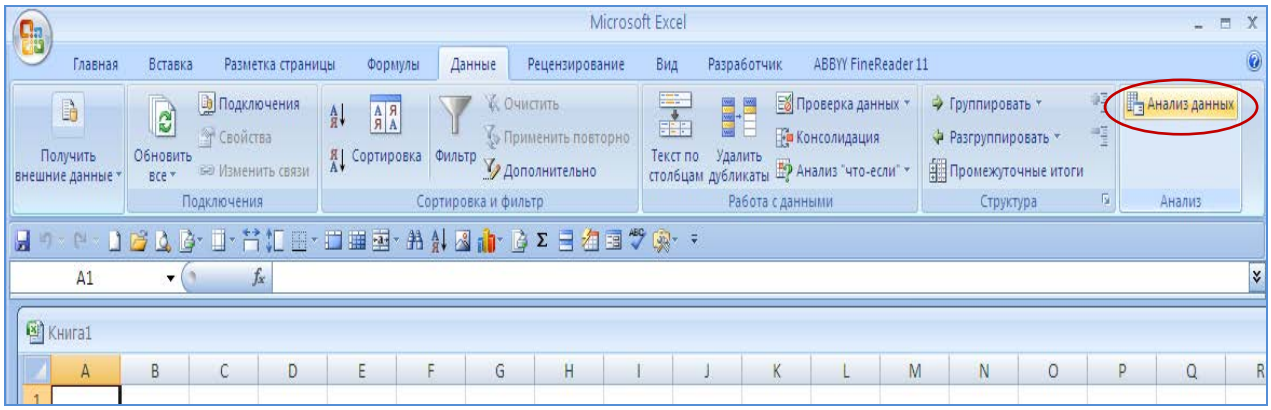


Рис. 7. Вид меню *Данные*

После вызова на выполнение пакета Анализ данных появится одноименное окно (рис. 8).

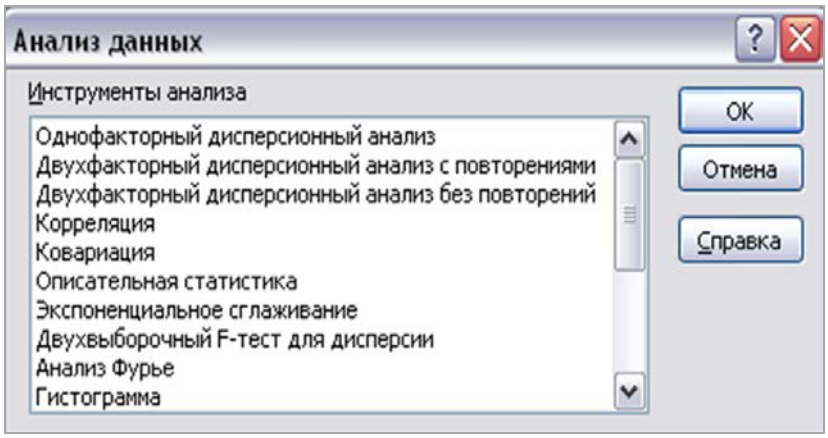


Рис. 8. Вид окна Анализ данных

Главным элементом этого окна является область ***Инструменты анализа.*** В данной области представлен список реализованных в пакете Анализ данных Microsoft Excel методов статистической обработки:

- «Гистограмма»;
- «Выборка»;
- «Описательная статистика»;

- «Ранг и персентиль»;
- «Генерация случайных чисел»;
- «Парный двухвыборочный *t*-тест для средних»;
- «Двухвыборочный *t*-тест с одинаковыми дисперсиями»;
- «Двухвыборочный *t*-тест с различными дисперсиями»;
- «Двухвыборочный *F*-тест для дисперсий»;
- «Двухвыборочный *z*-тест для средних»;
- «Однофакторный дисперсионный анализ»;
- «Двухфакторный дисперсионный анализ без повторений»;
- «Двухфакторный дисперсионный анализ с повторениями»;
- «Ковариация»;
- «Корреляция»;
- «Регрессия»;
- «Скользящее среднее»;
- «Экспоненциальное сглаживание»;
- «Анализ Фурье».

Каждый из перечисленных методов реализован в виде отдельного инструмента — процедуры. Для активизации соответствующей процедуры, необходимо ее выделить указателем мыши и щелкнуть по кнопке ОК.

Диалоговое окно каждой процедуры включает в себя элементы управления (поля ввода, раскрывающиеся списки, флажки, переключатели и т.п.), которые задают определенные параметры ее выполнения (в качестве примера на рисунке 9 изображено диалоговое окно процедуры «Гистограмма»).

Одна часть параметров является специфической и присуща только одному (или малой группе) инструментов. Назначение таких параметров будет рассмотрено при изучении технологий работы с соответствующими процедурами.

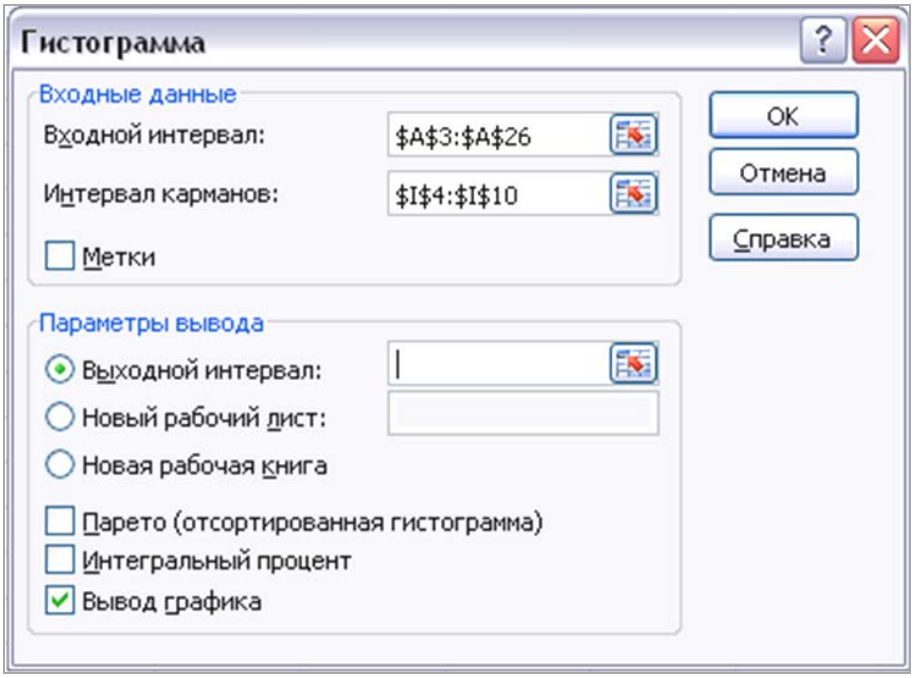


Рис. 9. Вид окна процедуры «Гистограмма»

Другая часть параметров универсальна и присуща всем (или подавляющему большинству) режимам работы. Элементами управления, задающими такие параметры, являются:

1. Поле *Входной интервал* — вводится ссылка на ячейки, содержащие анализируемые данные.

2. Переключатель *Группирование* — устанавливается в положение *По столбцам* или *По строкам* в зависимости от расположения данных во входном диапазоне.

3. Флажок *Метки* — устанавливается в активное состояние, если первая строка (столбец) во входном диапазоне содержит заголовки. Если заголовки отсутствуют, флажок не следует устанавливать. В этом случае будут автоматически созданы стандартные названия для данных выходного диапазона.

4. Переключатель *Выходной интервал/Новый рабочий лист/Новая рабочая книга.*

В положении *Выходной интервал* активизируется поле, в которое необходимо ввести ссылку на левую верхнюю ячейку выходного диапазона. Размер выходного диапазона будет определен автоматически, в случае возможного наложения выходного диапазона на исходные данные на экране появится сообщение.

В положении *Новый рабочий лист* открывается новый лист, в который, начиная с ячейки A1, вставляются результаты анализа. Если необходимо задать имя открываемого нового рабочего листа, введите его имя в поле, расположенное напротив соответствующего положения переключателя.

В положении *Новая рабочая книга* открывается новая книга, на первом листе которой, начиная с ячейки A1, вставляются результаты анализа.

Классификация показателей описательной статистики

Статистическая информация представляется совокупностью данных, для характеристики которых используются разнообразные показатели, называемые показателями *описательной статистики.* Показатели описательной статистики можно разбить на несколько групп [13].

1. ***Показатели положения*** описывают положение данных на числовой оси. Примеры таких показателей — минимальный и максимальный элементы выборки (первый и последний члены вариационного ряда), верхний и нижний квартили (ограничивают зону, в которую попадают 50% центральных элементов выборки). Сведения о середине совокупности могут дать среднее арифметическое, среднее гармоническое, медиана и другие характеристики.

2. ***Показатели разброса*** описывают степень разброса данных относительно своего центра. К ним относятся: дисперсия, стандартное отклонение, размах выборки (разность между максимальным и минимальным элементами), межквартильный размах (разность между верхним и нижним квартилем), эксцесс и т. п. Эти показатели определяют, насколько кучно ос-

новная масса данных группируется около центра.

3. ***Показатели асимметрии*** характеризуют симметрию распределения данных около своего центра. К ним можно отнести коэффициент асимметрии, положение медианы относительно среднего и т. п.

В пакете Анализ данных также есть ***показатели, описывающие закон распределения,*** которые дают представление о законе распределения данных. Сюда относятся таблицы частот, полигоны, кумуляты, гистограммы.

На практике чаще всего используются следующие показатели: среднее арифметическое, медиана, дисперсия, стандартное отклонение. Однако для получения более точных и достоверных выводов необходимо учитывать и другие из перечисленных выше характеристик, а также обращать внимание на условия получения выборочных совокупностей. Наличие выбросов, т.е. грубых ошибочных наблюдений, может не только сильно исказить значения выборочных показателей (выборочного среднего, дисперсии, стандартного отклонения и т. д.), но и привести ко многим другим ошибочным выводам.

Процедура «Описательная статистика» пакета Анализ данных

Процедура «Описательная статистика» служит для генерации одномерного статистического отчета по основным показателям положения, разброса и асимметрии выборочной совокупности.

В диалоговом окне данного режима (рис. 10) задаются следующие параметры:

1. Входной интервал.
2. Группирование.
3. Метки в первой строке/Метки в первом столбце.
4. Выходной интервал/Новый рабочий лист/Новая рабочая книга.
5. Итоговая статистика.
6. Уровень надежности и др.

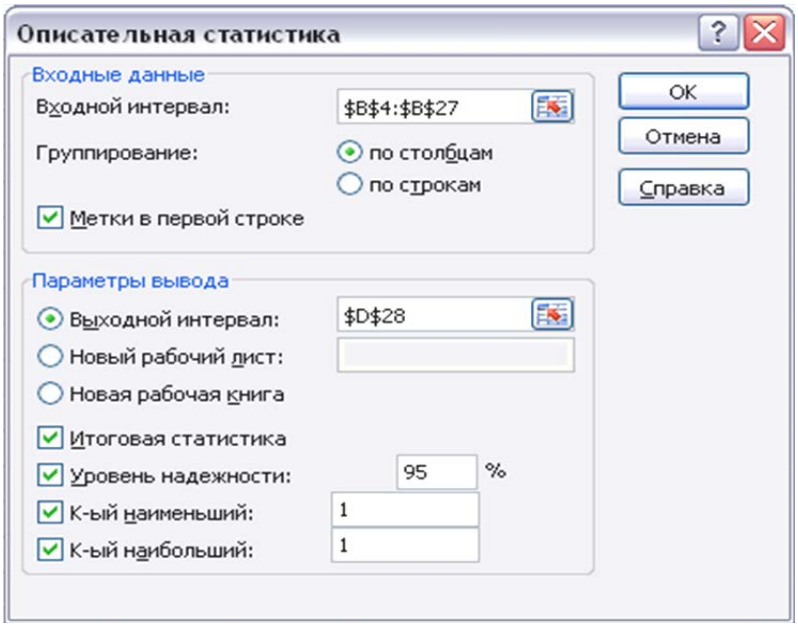


Рис. 10. Вид окна процедуры «Описательная статистика»

Установите флажок *Итоговая статистика* в активное состояние.

Уровень надежности — установите в активное состояние, если в выходную таблицу необходимо включить строку для предельной ошибки выборки ($\Delta x_{пр}$) — (предельной ошибки среднего) при установленном уровне надежности. В поле, расположенном напротив флажка, введите требуемое значение уровня надежности (например, значение уровня надежности 95% равносильно доверительной вероятности γ=0,95 или уровню значимости α = 0,05).

К-й наибольший — установите в активное состояние, если в выходную таблицу необходимо включить строку для k-го наибольшего (начиная с максимума X_{max}) значения элемента выборки. В поле, расположенное напротив флажка, введите число *k*. Если *k*=1, то строка будет содержать максимальное значение элемента выборки.

К-й наименьший — установите в активное состояние, если в выходную таблицу необходимо включить строку для k-го наименьшего (начиная с минимума X_{min}) значения элемента выборки. В поле, расположенное напротив флажка, введите число *k*. Если *k* = 1, то строка будет содержать минимальное значение элемента выборки.

Процедура «Гистограмма» пакета Анализ данных

Процедура «Гистограмма» служит для вычисления частот попадания данных в указанные границы интервалов, а также для построения гистограммы *интервального* вариационного ряда. Вид окна указанной процедуры представлен на рисунке 11.

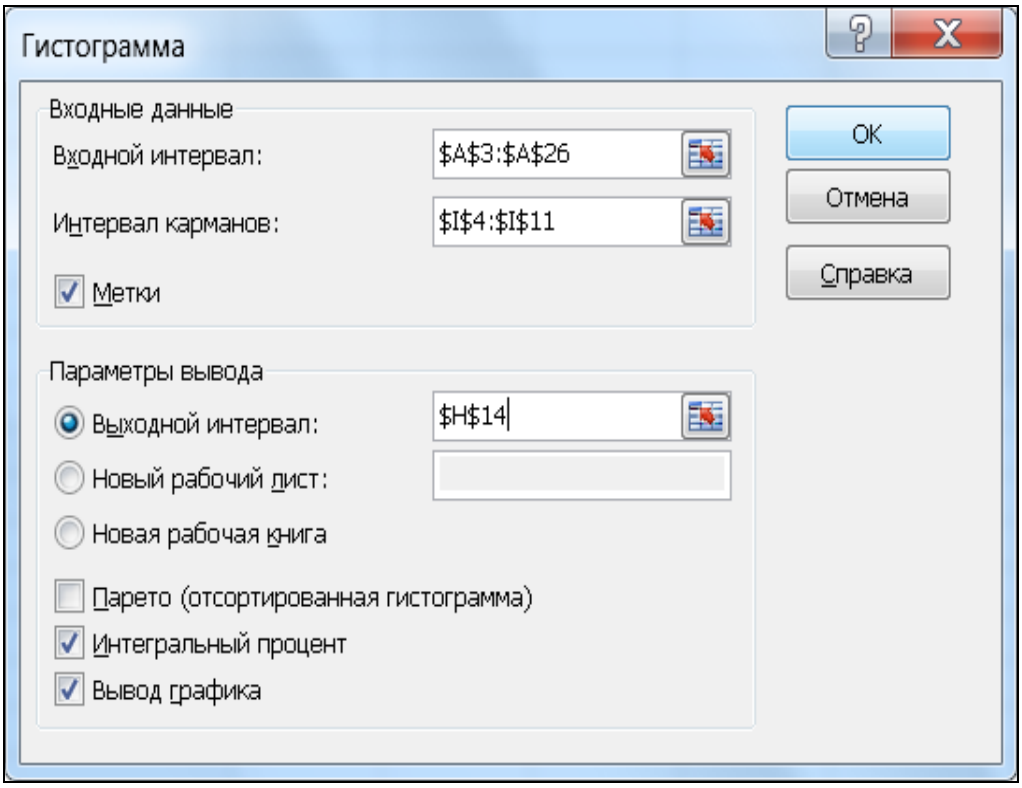


Рис. 11. Вид окна процедуры «Гистограмма»

В диалоговом окне данной процедуры задаются следующие параметры:

- *Входной интервал.*
- *Интервал карманов* (необязательный параметр) — вводится ссылка на ячейки, содержащие набор граничных значений, определяющих интервалы (карманы). Эти значения должны быть введены в возрастающем порядке. В Microsoft Excel вычисляется число попаданий данных в сформированные интервалы, причем границы интервалов являются строгими нижними границами и нестрогими верхними: $a<x\leq b$

Если диапазон карманов не был введен, то набор интервалов, равномерно распределенных между минимальным и максимальным значениями данных, будет создан автоматически.

- *Метки.*
- *Выходной интервал/Новый рабочий лист/Новая рабочая книга.*
- *Парето (отсортированная гистограмма*) — устанавливается в активное состояние, чтобы представить данные в порядке убывания частоты. Если флажок снят, то данные в выходном диапазоне будут приведены в порядке следования интервалов.
- *Интегральный процент* — устанавливается в активное состояние для расчета выраженных в процентах *накопленных* частот и включения его в гистограмму.
- *Вывод графика* — устанавливается в активное состояние для автоматического создания встроенной диаграммы на листе, содержащем входной диапазон.

В режиме работы «Гистограмма» пользователь может самостоятельно задать величину интервалов ряда (параметр *Интервал карманов* диалогового окна Гистограмма), а может не задавать, тогда эти значения будут вычислены автоматически.

3.2. ВОЗМОЖНОСТИ ПАКЕТА АНАЛИЗ ДАННЫХ. ОПИСАТЕЛЬНАЯ СТАТИСТИКА, ГИСТОГРАММА ЧАСТОТНОГО РАСПРЕДЕЛЕНИЯ ВЫБОРКИ

ЦЕЛИ ЗАНЯТИЯ

1. Научиться вычислять описательную статистику выборки с помощью процедуры «Описательная статистика» пакета Анализ данных.

2. Научиться формулировать выводы о нормальности частотного распределения выборки по полученным результатам.

3. Овладеть навыками применения процедуры «Гистограмма» пакета Анализ данных для построения гистограммы частотного распределения выборки.

МЕТОДИКА ВЫПОЛНЕНИЯ РАБОТЫ

Постановка задачи 1

Дано: Выборочные совокупности, содержащие сведения о температуре пациентов двух групп: контрольной и исследуемой.

Контрольная	*Исследуемая*
38,5	37,6
38,2	37,1
39	38,1
39,5	38,2
38,7	38
37	37,9
38,4	36,8
38,3	37,1
39,2	38,2
37,9	36,8
37,9	36,5
38,4	38,1
38,6	38,2
38,4	36,7
37,3	36,9
37,7	36,8
37,4	37
37	37,5
38,5	37,6
37,9	37,2
37,8	37,3
38	37,6
38,8	38,2

Требуется:

1. Вычислить основные характеристики двух выборок с помощью встроенных функций.

2. Используя один из известных критериев нормальности частотного распределения, сделать вывод о нормальности частотного распределения в группах.

3. Обосновать и записать в тетрадь выбор методов дальнейшей статистической обработки данных (параметрический, непараметрический).

4. Построить с помощью Мастера диаграмм программы Microsoft Excel гистограммы частотного распределения выборок, с заданными пользователем диапазонами признака.

Постановка задачи 2

Дано: Результаты задачи 1.

Требуется:

1. На листе книги Microsoft Excel с рассчитанными описательными статистиками для двух групп получить аналогичные данные с помощью процедуры пакета Анализ данных сначала для контрольной группы, а затем для исследуемой.
2. По полученным данным рассчитать доверительный интервал (ДИ) для среднего.
3. С помощью процедуры пакета Анализ данных получить гистограммы последовательно для каждой из выборок.

ХОД РАБОТЫ

Решение задачи 1

1. Скопируйте из папки «Z:\ Материалы для работы\Статистика» в свою папку файл *Пр.зан.№2-Гистограмма.xls*.
2. Измените имя файла на *«Пр.зан.№2- <своя фамилия группа>.xls»*.
   Например, *Пр.зан.№2- Иванов А. — 24 леч>.xls.*
3. Вычислите основные характеристики двух выборок с помощью встроенных функций Microsoft Excel.
4. Используя один из известных критериев нормальности частотного распределения, сделайте и запишите в тетрадь вывод о нормальности частотных распределений в группах.
5. Обоснуйте выбор методов дальнейшей статистической обработки данных (параметрический, непараметрический), запишите обоснование в тетрадь.
6. На Рабочем листе с именем *«Гистограмма»* постройте с помощью Мастера диаграмм гистограмму частотного распределения выборок, предварительно вычислив частоты встречаемости признака в каждой группе по указанным ниже диапазонам. Занесите данные в тетрадь в таблицу 3.

Таблица 3.

Частоты встречаемости признака в группах

Диапазоны значений	**Частота**	
	Контрольная	***Исследуемая***
36,5 -37		
37,1 - 37,5		
37,6 - 38		
38,1 - 38,5		
38,6 - 39		
39,1 -39,5		
39,6- 40		

Общий вид Рабочего листа *«Гистограмма»* с результатами, представлен на рисунке 12.

	A	B	C	D	E	F	G
1	Построение гистограммы						
2	*Группы*						
3	*Контрольная*	*Исследуемая*					
4	37	36,5		Диапазоны	Частота		
5	37	36,7		значений	*Контрольная*	*Исследуемая*	
6	37,3	36,8		36,5 -37	2	7	
7	37,4	36,8		37,1 - 37,5	2	5	
8	37,7	36,8		37,6 - 38	6	5	
9	37,8	36,9		38,1 - 38,5	6	6	
10	37,9	37		38,6 - 39	4	0	
11	37,9	37,1		39,1 -39,5	3	0	
12	37,9	37,1		39,6- 40	0	0	
13	38	37,2					
14	38,2	37,3					
15	38,3	37,5					
16	38,4	37,6					
17	38,4	37,6					
18	38,4	37,6					
19	38,5	37,9					
20	38,5	38					
21	38,6	38,1					
22	38,7	38,1					
23	38,8	38,2					
24	39	38,2					
25	39,2	38,2					
26	39,5	38,2					
27							
28							
29							
30							
31							

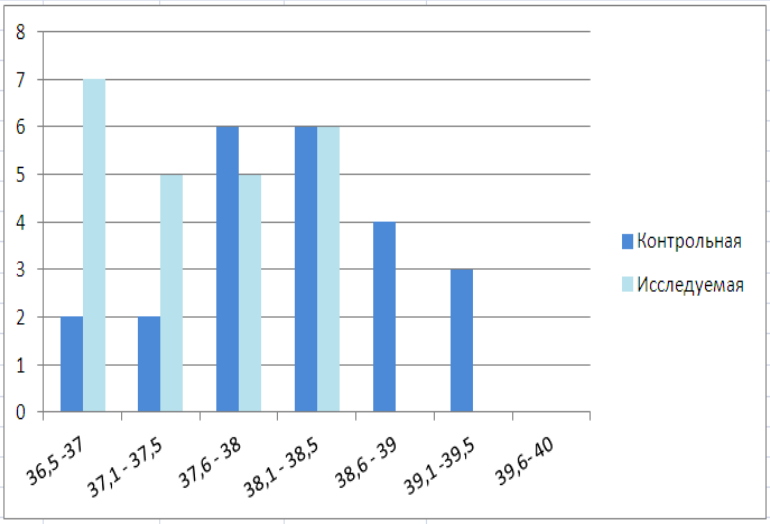


Рис. 12. Общий вид Рабочего листа *«Гистограмма»*

✓ **Обратите внимание!** На рисунке12. и последующих для визуализации материала приведен только общий вид результатов. Данные в примерах могут отличаться от тех, что заданы в практической работе.

Решение задачи 2

Процедура пакета Анализ данных «Описательная статистика»

1. На листе «*Задание 1*» рабочей книги Microsoft Excel с рассчитанными описательными статистиками получите аналогичные данные с помощью процедуры «Описательная статистика» пакета Анализ данных.

Для этого:

- вызовите пакет Анализ данных, применив команду: ***Данные⇒ Анализ данных;***
- в окне «Анализ данных» (рис. 13) активируйте процедуру «***Описательная статистика***»;

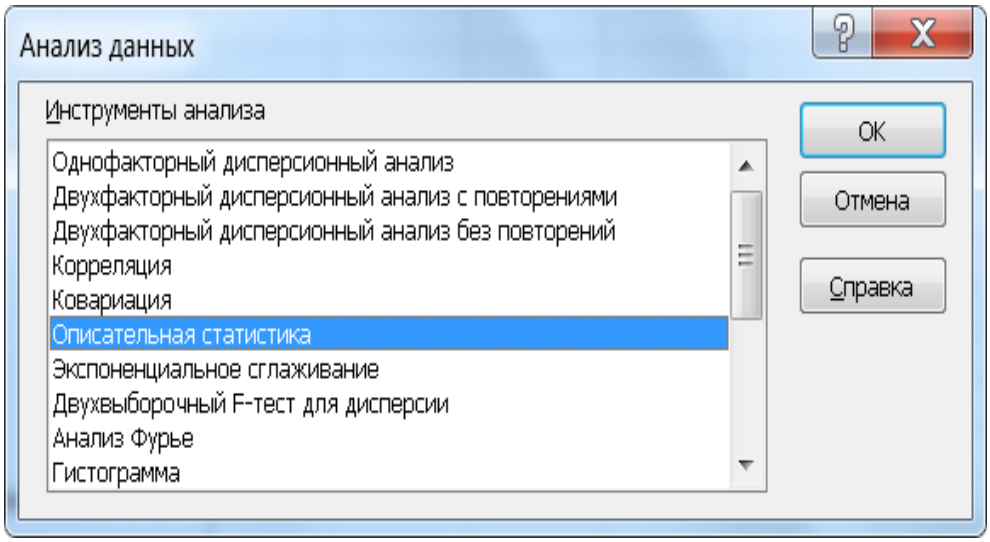


Рис. 13 Вид окна Анализ данных

- в соответствии с рисунком 14 установите флажки и заполните нужные поля для вычисления описательной статистики для ***контрольной*** группы;
- нажмите ОК.

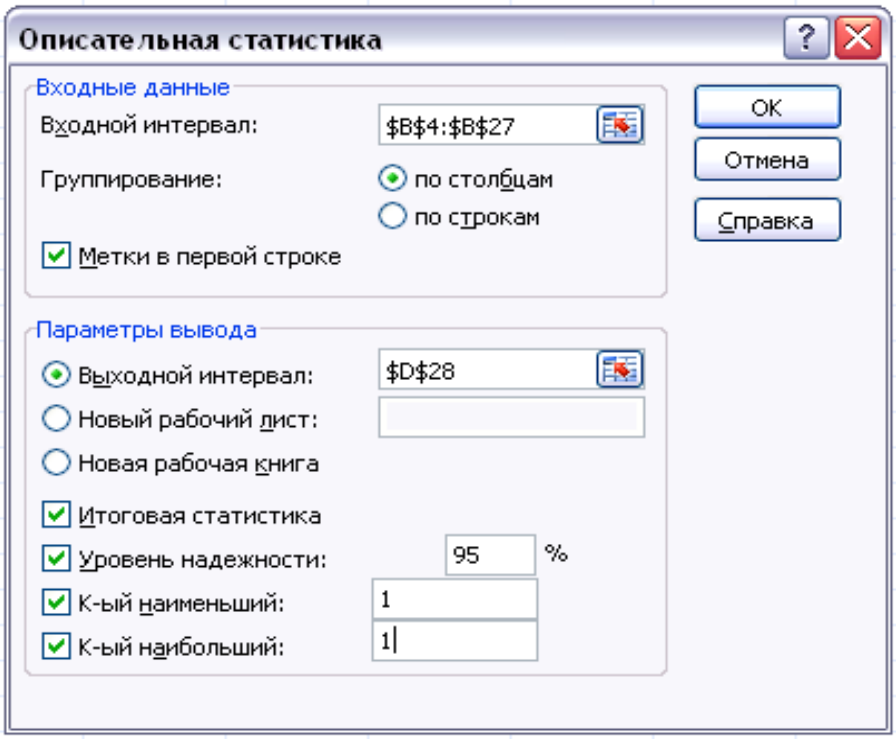
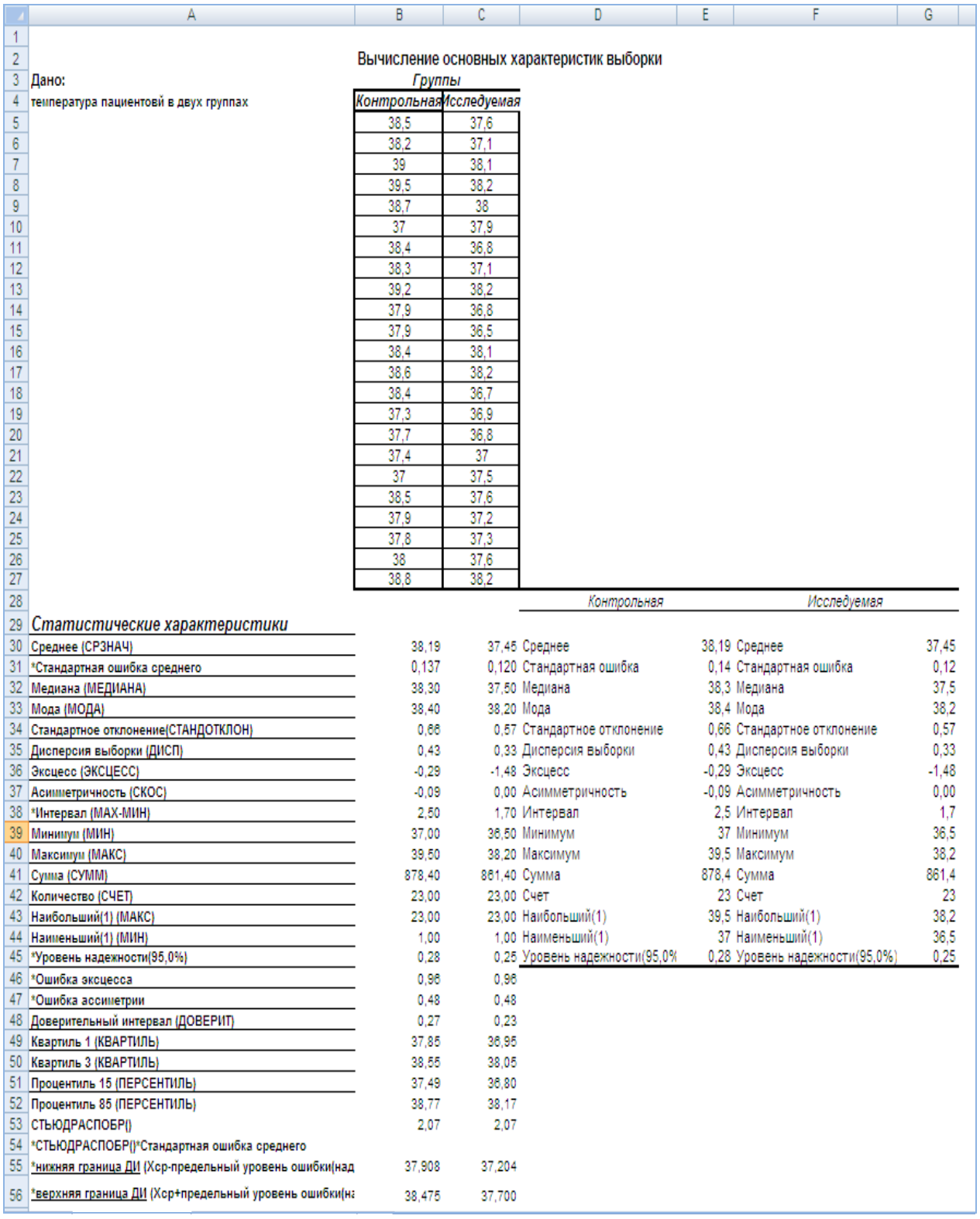


Рис. 14. Вид окна «Описательная статистика» с введенными параметрами для контрольной группы

- Повторите выполнение процедуры «Описательная статистика» для ***исследуемой*** группы, изменив значение полей: Входной интервал и Выходной интервал ($F$28).

Общий вид результатов на Рабочем листе представлен на рисунке 15.



Вычисление основных характеристик выборки

Дано:

температура пациентовй в двух группах

Группы	
Контрольная	Исследуемая
38,5	37,6
38,2	37,1
39	38,1
39,5	38,2
38,7	38
37	37,9
38,4	36,8
38,3	37,1
39,2	38,2
37,9	36,8
37,9	36,5
38,4	38,1
38,6	38,2
38,4	36,7
37,3	36,9
37,7	36,8
37,4	37
37	37,5
38,5	37,6
37,9	37,2
37,8	37,3
38	37,6
38,8	38,2

Статистические характеристики						
			Контрольная		Исследуемая	
Среднее (СРЗНАЧ)	38,19	37,45	Среднее	38,19	Среднее	37,45
*Стандартная ошибка среднего	0,137	0,120	Стандартная ошибка	0,14	Стандартная ошибка	0,12
Медиана (МЕДИАНА)	38,30	37,50	Медиана	38,3	Медиана	37,5
Мода (МОДА)	38,40	38,20	Мода	38,4	Мода	38,2
Стандартное отклонение(СТАНДОТКЛОН)	0,66	0,57	Стандартное отклонение	0,66	Стандартное отклонение	0,57
Дисперсия выборки (ДИСП)	0,43	0,33	Дисперсия выборки	0,43	Дисперсия выборки	0,33
Эксцесс (ЭКСЦЕСС)	-0,29	-1,48	Эксцесс	-0,29	Эксцесс	-1,48
Асимметричность (СКОС)	-0,09	0,00	Асимметричность	-0,09	Асимметричность	0,00
*Интервал (МАХ-МИН)	2,50	1,70	Интервал	2,5	Интервал	1,7
Минимум (МИН)	37,00	36,50	Минимум	37	Минимум	36,5
Максимум (МАКС)	39,50	38,20	Максимум	39,5	Максимум	38,2
Сумма (СУММ)	878,40	861,40	Сумма	878,4	Сумма	861,4
Количество (СЧЕТ)	23,00	23,00	Счет	23	Счет	23
Наибольший(1) (МАКС)	23,00	23,00	Наибольший(1)	39,5	Наибольший(1)	38,2
Наименьший(1) (МИН)	1,00	1,00	Наименьший(1)	37	Наименьший(1)	36,5
*Уровень надежности(95,0%)	0,28	0,25	Уровень надежности(95,0%	0,28	Уровень надежности(95,0%)	0,25
*Ошибка эксцесса	0,96	0,96				
*Ошибка ассиметрии	0,48	0,48				
Доверительный интервал (ДОВЕРИТ)	0,27	0,23				
Квартиль 1 (КВАРТИЛЬ)	37,85	36,95				
Квартиль 3 (КВАРТИЛЬ)	38,55	38,05				
Процентиль 15 (ПЕРСЕНТИЛЬ)	37,49	36,80				
Процентиль 85 (ПЕРСЕНТИЛЬ)	38,77	38,17				
СТЬЮДРАСПОБР()	2,07	2,07				
*СТЬЮДРАСПОБР()*Стандартная ошибка среднего						
*нижняя граница ДИ (Хср-предельный уровень ошибки(над	37,908	37,204				
*верхняя граница ДИ (Хср+предельный уровень ошибки(на	38,475	37,700				

Рис. 15. Вид Рабочего листа с результатами выполнения задания

2. Запишите результаты в тетрадь в таблицу 4.

3. Вычислите и запишите в тетрадь (таблица 4) значения верхней и нижней границ доверительного интервала для каждой группы:

Нижняя граница ДИ = Среднее - Уровень надежности (95%);
Верхняя граница ДИ = Среднее + Уровень надежности (95%).

Таблица 4.

Описательные статистики двух групп

Статистические характеристики	Группа	
	Контрольная	Исследуемая
Среднее		
Стандартная ошибка		
Медиана		
Мода		
Стандартное отклонение		
Дисперсия выборки		
Эксцесс		
Асимметричность		
Интервал		
Минимум		
Максимум		
Сумма		
Счет		
Наибольший(1)		
Наименьший(1)		
Уровень надежности(95,0%)		
Нижняя граница ДИ		
Верхняя граница ДИ		

4. Сравните данные, полученные с помощью процедуры «Описательная статистика» с данными, полученными с помощью встроенных функций Microsoft Excel.

5. Запишите в тетрадь вывод о том, как отличаются результаты, полученные разными методами. Какой метод более рационален?

6. Сохраните результаты работы в Вашем файле.

Процедура «Гистограмма» пакета Анализ данных

1. Перейдите на Рабочий лист с именем «*Гистограмма*» в Вашем файле книги Microsoft Excel.

2. Подготовьте на листе *«Гистограмма»* интервал «карманов» (концы диапазонов), заполнив в таблице Microsoft Excel с ячейки ***I4*** следующий столбец значений, если его нет на Рабочем листе.

температура
37
37,5
38
38,5
39
39,5
40

3. Вызовите процедуру «Гистограмма» из пакета Анализ данных.

4. Для построения гистограммы частотного распределения **кон-**

трольной группы заполните значения соответствующих полей и установите флажки: *Интегральный процент* и *Вывод графика* в соответствии с рисунком 16.

5. Нажмите **ОК**.

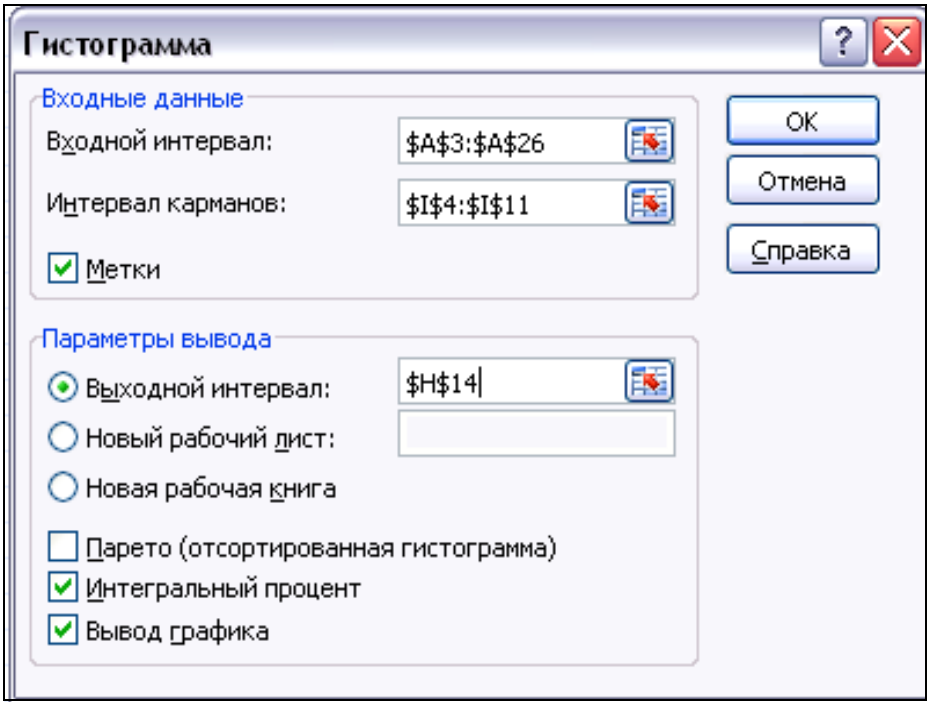


Рис. 16. Вид окна процедуры «Гистограмма» для получения гистограммы частотного распределения контрольной группы

6. Выведите гистограмму для *исследуемой* группы с ячейки ***H25***.

7. Разместите гистограммы на листе *«Гистограмма»* в соответствии с рисунком 17.

8. Сравните построенную Вами гистограмму с помощью Мастера диаграмм с гистограммами, полученными с помощью аналогичной процедуры пакета Анализ данных.

9. Сделайте вывод, о том какой способ построения гистограмм частотных распределений является более рациональным, запишите его в тетрадь.

ДОПОЛНИТЕЛЬНОЕ ЗАДАНИЕ

1. Под результатами практической работы на Рабочем листе *«Гистограмма»* файла Microsoft Excel самостоятельно получите гистограммы частотных распределений для двух выборок, не задавая интервал карманов, т.е. применив его значение «по умолчанию».

2. Сравните полученные гистограммы частот с применением диапазона карманов с теми, что созданы без их применения, т.е. автоматически.

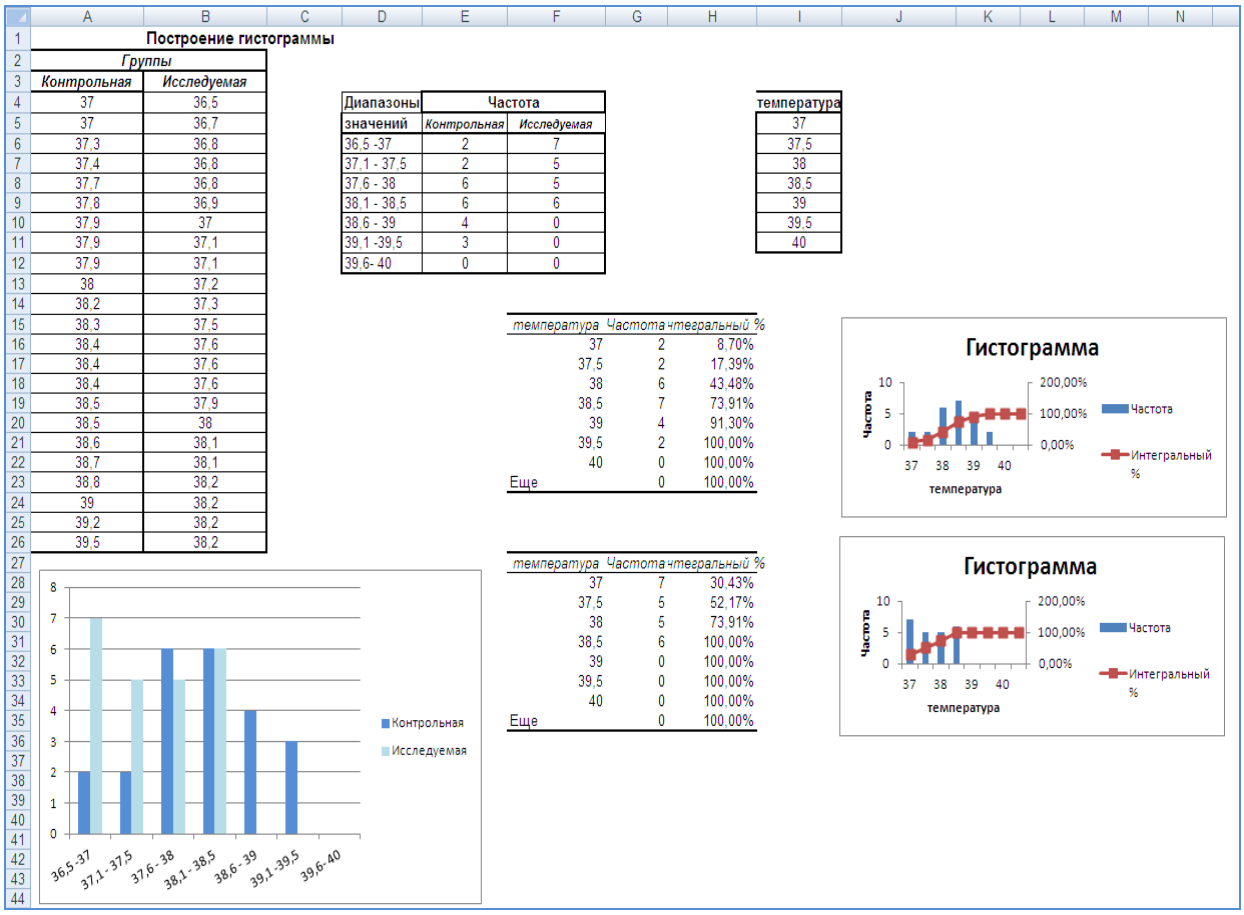


Рис. 17. Общий вид листа *«Гистограмма»* с результатами работы

3.3. ПРОЦЕДУРЫ ПАКЕТА АНАЛИЗ ДАННЫХ, ОСУЩЕСТВЛЯЮЩИЕ ФОРМИРОВАНИЕ И АНАЛИЗ ОСНОВНЫХ ХАРАКТЕРИСТИК ВЫБОРКИ

Пакет Анализа данных Microsoft Excel включает ряд процедур (инструментов), не осуществляющих определенный вид статистического анализа данных, но позволяющих выполнять ряд полезных операций при формировании репрезентативной выборки, анализе ее свойств и структуры. К ним следует отнести следующие процедуры.

Генерация случайных чисел — генерирует псевдослучайные числа, соответствующие заданному пользователем закону частотного распределения, позволяет создать рандомизированную выборку.

Выборка — позволяет случайным образом из большого массива данных (генеральной совокупности) сформировать (выбрать) выборку с заданным количеством вариант, которую можно в дальнейшем использовать в медико-биологическом исследовании.

Ранг и персентиль — осуществляет ***ранжирование*** вариант в выборке: каждому объекту выборки в соответствии с его значением присваивает определенный ранг (большему значению меньший ранг) и процент, характеризующий его место в частотном распределении выборки. Эти данные могут быть полезны при реализации непараметрических методов анализа,

например, при вычислении значения коэффициента ранговой корреляции Спирмена.

Гистограмма — выводит гистограмму частотного распределения выборки. Наличие гистограммы позволяет визуализировать процесс обработки данных.

Подходы к формированию рандомизированной выборки

Методология исследования массовых статистических явлений в зависимости от полноты охвата изучаемого объекта (явления) различает *сплошное* и *несплошное* наблюдение. Разновидностью несплошного наблюдения является выборочное.

Под *выборочным* наблюдением понимается метод статистического исследования, при котором обобщающие показатели изучаемой совокупности устанавливаются по некоторой ее части на основе случайного отбора. При выборочном методе обследованию подвергается сравнительно небольшая часть всей изучаемой совокупности, получившая название *выборочной совокупности* или просто *выборки.*

Выборка должна быть *представительной (репрезентативной),* чтобы по ней можно было судить о генеральной совокупности. Репрезентативность означает, что объекты выборки достаточно хорошо представляют генеральную совокупность.

Выборка описывается рядом параметров, среди которых: закон частотного распределения элементов в выборке, среднее арифметическое, медиана, мода, дисперсия, стандартное отклонение, максимальное и минимальное значения, асимметрия, эксцесс и другие величины.

Элементы выборки обычно распределяются в соответствии с каким-то законом, который чаще всего можно описать математически (биномиальное, распределение Пуассона, нормальное распределение, распределение Фишера и др.). Для исследователя является принципиальным, подчиняются ли элементы выборки нормальному закону частотного распределения — закону Гаусса или какому-то другому.

Одним из способов предупреждения систематических ошибок выборочного исследования является формирование собственно-случайной выборочной совокупности.

Собственно-случайная выборка состоит в том, что выборочная совокупность образуется в результате случайного (непреднамеренного) отбора отдельных единиц из генеральной совокупности.

На практике, особенно в генеральной совокупности большого объема, для организации собственно-случайной выборки часто используют таблицу случайных чисел или генератор случайных чисел. В Microsoft Excel в одном из способов формирования выборки также используется ***генератор случайных чисел.***

Процедура «Генерация случайных чисел»

Одним из фундаментальных в статистическом анализе является понятие *случайной величины.* Случайной называется переменная величина, принимающая в зависимости от случая те или иные значения с определенными вероятностями. В практических задачах обычно используются **дискретные** и **непрерывные** случайные величины.

Дискретной случайной величиной называется такая величина, множество возможных значений которой можно посчитать. Значения этой величины являются целыми положительными числами. Например, количество родов у женщины.

Непрерывной случайной величиной называется такая случайная величина, которая может принять любое значение из некоторого конечного или бесконечного интервала. Например, температура пациента.

Чтобы дать полное математическое описание случайной величины, нужно указать множество ее значений и соответствующее этой величине распределение вероятностей на исследуемом множестве.

В статистике широко используются различные виды теоретических распределений — нормальное распределение, биномиальное, распределение Пуассона и др. Каждое из теоретических распределений имеет специфику и свою область применения. Чаще всего в качестве теоретического распределения используется *нормальное распределение,* занимающее особое положение в статистических исследованиях.

Справочная информация по технологии работы

Процедура «Генерация случайных чисел» служит для формирования массива случайных чисел, распределенных по одному из заданных теоретических распределений, и может использоваться для получения репрезентативной выборки, сформированной случайным образом. Такая выборка называется рандомизированной.

В зависимости от выбранного теоретического распределения (режима работы) меняются и параметры диалогового окна процедуры «***Генерация случайных чисел».*** Общими параметрами для всех режимов являются следующие.

1. ***Число переменных*** — вводится число **столбцов** значений, которые необходимо разместить в выходном диапазоне. Если это число не введено, то все столбцы в выходном диапазоне будут заполнены.

2. ***Число случайных чисел*** — вводится число случайных значений, которое необходимо вывести в каждом столбце выходного диапазона. Каждое случайное значение будет помещено в строке выходного диапазона. Если число случайных чисел не будет введено, все строки выходного диапазона будут заполнены.

3. ***Распределение*** — в данном раскрывающемся списке выбирается тип распределения, которое необходимо использовать для генерации слу-

чайных чисел.

4. ***Случайное рассеивание*** — вводится «стартовое» число для генерации определенной последовательности случайных чисел. Впоследствии это число можно снова использовать для получения той же самой последовательности случайных чисел.

5. ***Выходной интервал/Новый рабочий лист/Новая рабочая книга*** — указывают адрес первой ячейки для размещения результата.

Технология работы во всех режимах процедуры «Генерация случайных чисел» является одинаковой, отличается только задание значений в области *Параметры*, характерных для конкретных распределений.

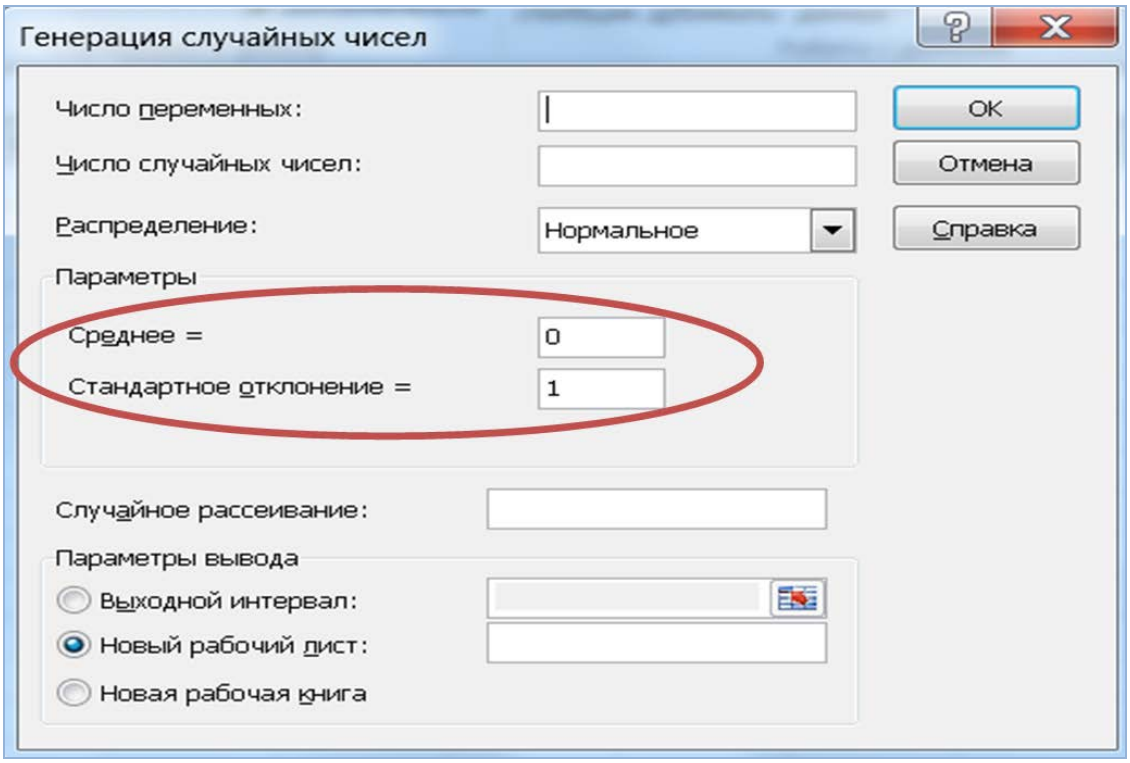


Рис. 18. Вид диалогового окна процедуры «Генерация случайных чисел»

На рисунке 18 представлено диалоговое окно режима работы, предназначенного для генерации случайных чисел, распределенных по *нормальному* закону. В этом окне в области *Параметры* задаются характеристики нормального закона распределения — математическое ожидание (поле «*Среднее»)* и стандартное отклонение (поле «*Стандартное отклонение»).*

Для генерации последовательности случайных чисел, распределенных по ***биномиальному*** закону, в области *Параметры* задаются вероятность успеха при одном испытании (поле *Значение p)* и число испытаний (поле *Число испытаний*).

Пример применения процедуры

Дано: Номера историй болезни пациентов с конкретной патологией, которых лечат в данном отделении больницы, находятся в диапазоне от 305 до 420.

Требуется: С помощью генератора случайных чисел сгенерировать последовательность (**30**) номеров историй болезни пациентов, данные которых будут использованы при формировании рандомизированной выборки.

Решение задачи

Из пакета Анализ данных вызовем процедуру «Генерация случайных чисел» и заполним поля ее диалогового окна в соответствии с рисунком 19.

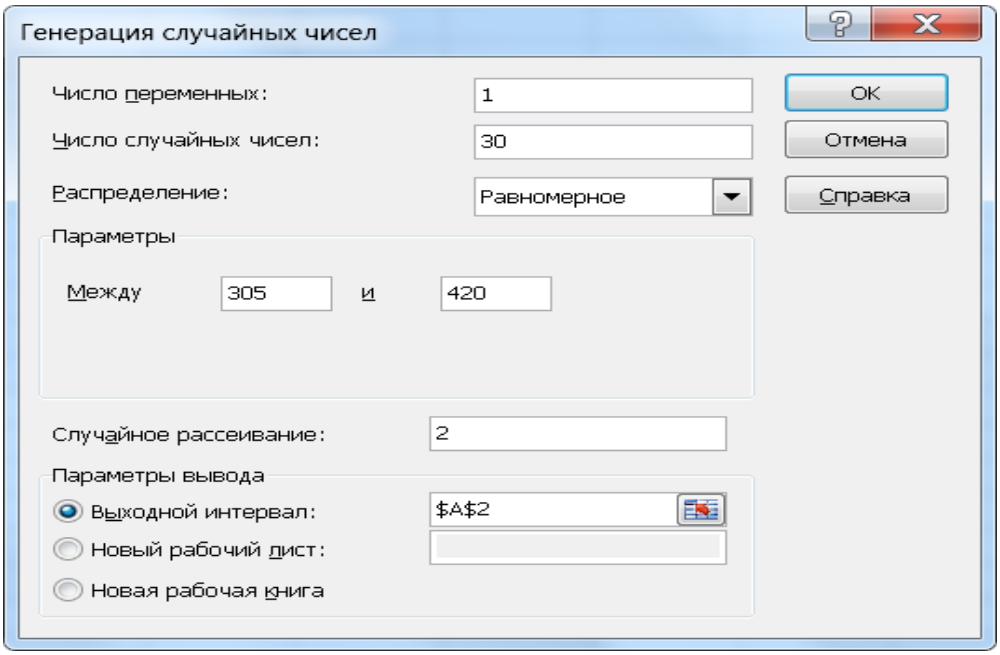
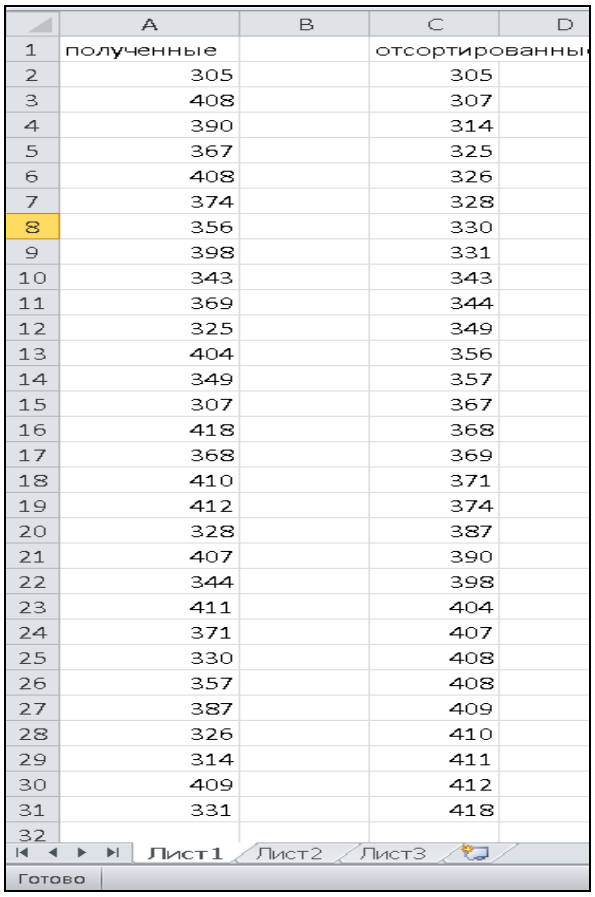


Рис. 19. Вид диалогового окна процедуры «Генерация случайных чисел» с введенными параметрами

Так как о частотном распределении ничего не сказано, используем *равномерное распределение. Число переменных* (столбцов) равно ***1***, требуемое *число случайных чисел* — ***30***. *Случайное рассеивание* (стартовое число) — это произвольное число, которое всегда будет обеспечивать нам вывод одной и той же последовательности чисел. Введем произвольное значение ***2***. Для того чтобы повторить вывод этой последовательности, стартовое число надо запомнить.

Округлим числа, полученные в результате работы процедуры, до целых, использовав пиктограмму «*Уменьшить разрядность*». Для выявления повторов скопируем результаты в другую колонку электронной таблицы, упорядочим по возрастанию значений, удалим повторы, если они есть. Результаты приведены на рисунке 20. В сгенерированной последовательности обнаружили повторы значения 408. Одно из этих чисел отбросим и сгенерируем последовательность еще раз с тем же стартовым числом 2, но увеличим количество случайных чисел на 2.

Полученный результат округлим, уменьшив разрядность, отсортируем и опять удалим найденные повторы. Полученные значения и будут являться номерами историй болезни пациентов, данные которых мы должны использовать для формирования рандомизированной выборки.



	A	B	C	D
1	полученные		отсортированны	
2	305		305	
3	408		307	
4	390		314	
5	367		325	
6	408		326	
7	374		328	
8	356		330	
9	398		331	
10	343		343	
11	369		344	
12	325		349	
13	404		356	
14	349		357	
15	307		367	
16	418		368	
17	368		369	
18	410		371	
19	412		374	
20	328		387	
21	407		390	
22	344		398	
23	411		404	
24	371		407	
25	330		408	
26	357		408	
27	387		409	
28	326		410	
29	314		411	
30	409		412	
31	331		418	
32				

Лист1 Лист2 Лист3

Готово

Рис. 20. Результаты работы процедуры «Генерация случайных чисел»

Процедура «Выборка»

Процедура «Выборка» используется для реализации отбора с *заданным шагом* или *случайным образом* из большой совокупности части ее значений.

Предположим, что при лечении пациентов с определенной патологией, используя новую методику, были получены данные ряда поликлиник областного центра. Для подтверждения эффективности этой методики необходимо сформировать контрольную выборку из показателей 100 пациентов. Такая выборка является случайной с повторением, так как некоторые пациенты могут быть выбраны дважды. Если же необходимо организовать случайную выборку без повторения, то вновь встретившееся значение поля ФИО следует пропустить и осуществить выборку данных повторно.

При формировании *малых выборок* достаточно сложной проблемой является определение необходимого (оптимального) объема выборки. В математической статистике доказывается, что необходимая численность *собственно-случайной повторной* выборки определяется выражением:

$$n = \frac{t^2\sigma^2}{\Delta_x^2}$$

где n — количество объектов в выборке;

Δ_x — предельная ошибка выборки;
σ^2 — дисперсия генеральной совокупности;
t — коэффициент доверия (Стьюдента). Он определяется в зависимости от того, с какой доверительной вероятностью надо гарантировать результаты выборочного обследования.

Для оценки генеральной дисперсии σ^2 используют материалы предыдущих исследований, существующие нормативы, или проводят пробное обследование. По результатам пробного обследования оценивают значение генеральной дисперсии для последующего обоснования необходимого объема выборки.

Справочная информация по технологии работы

Процедура «Выборка» пакета Анализ данных предусматривает возможность отбора данных из предложенного массива двумя способами, которые представлены в разделе ***Метод выборки*** диалогового окна процедуры (рис. 21):

- ***периодический,*** при использовании этого метода задается ***период*** (шаг) выбора данных из предложенной последовательности;
- ***случайный*** (с использованием генератора случайных чисел), задается ***число выборок*** (количество выбранных чисел из исходного массива). В этом случае в результатах работы процедуры возможны повторы выбранных значений, которые следует удалить.

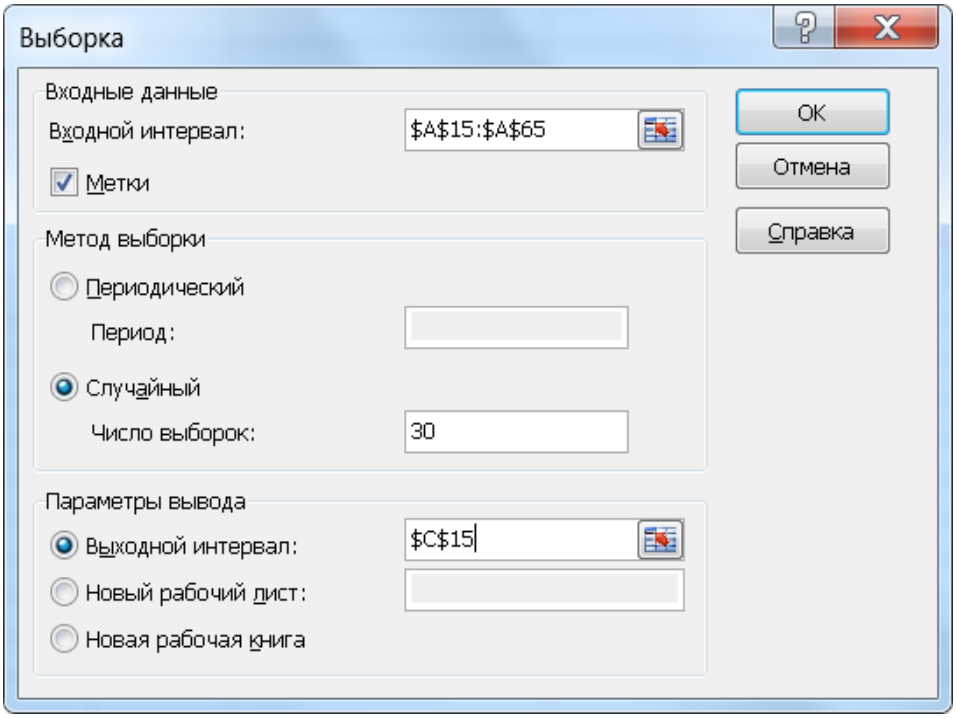


Рис. 21. Вид диалогового окна процедуры «Выборка»

Остальные поля диалогового окна заполняются, как и в других процедурах пакета Анализ данных.

Процедура «Ранг и персентиль»

При проведении анализа взаимного расположения значений признака в выборке наряду с такими понятиями, как мода, медиана, квартиль, квантиль, дециль, персентиль, пользуются также понятиями *ранга* и *процентранга.*

Под *рангом (R)* понимают номер (порядковое место) значения случайной величины в наборе данных. Правила присвоения рангов состоят в следующем:

1. если в наборе данных все числа разные, то каждому числу x_i присваивается уникальный ранг R_i;
2. если в наборе данных встречается группа из k одинаковых чисел $x_i = x_{i+1} = x_{i+2} = \ldots = x_{i+k-1}$, то ранг у них одинаковый и равен рангу первого числа из этой группы R_i. Число, следующее за этой группой, получает ранг, равный R_{i+k};
3. если данные упорядочены *по убыванию*, то:

   а) максимальное значение в наборе данных имеет ранг, равный 1;

   б) минимальное значение в наборе данных имеет наибольшее значение ранга, равное n-k_{min}+1, где n — количество данных в наборе, k_{min} — количество повторяющихся минимальных значений в наборе данных;
4. если данные упорядочены *по возрастанию*, то:

   а) минимальное значение в наборе данных имеет ранг, равный 1;

   б) максимальное значение в наборе данных имеет наибольшее значение ранга, равное n-k_{max}+1, где n — количество данных в наборе, k_{max} — количество повторяющихся максимальных значений в наборе данных.

Под *процентрангом* понимают процентное отношение для каждого значения в наборе данных.

Ранги характеризуют взаимное расположение значений признака в наборе данных, а также находят практическое применение в *непараметрических* методах оценки взаимосвязи медико-биологических, социально-экономических явлений и процессов.

В частности, ранги входят в формулу расчета коэффициента Спирмена, который может быть использован для определения силы связи, как между количественными, так и качественными признаками при условии, что их значения могут быть упорядочены по убыванию или возрастанию.

Справочная информация по технологии работы

Процедура «Ранг и персентиль» служит для генерации таблицы, содержащей порядковые и процентные ранги для каждого значения из набора данных, при этом данные упорядочиваются *по убыванию.*

В диалоговом окне данного режима (рис. 22) задаются параметры:

1. *Входной интервал.*
2. *Группирование (по столбцам или строкам).*
3. *Метки в первой строке/Метки в первом столбце.*

4. *Выходной интервал/Новый рабочий лист/Новая рабочая книга.*

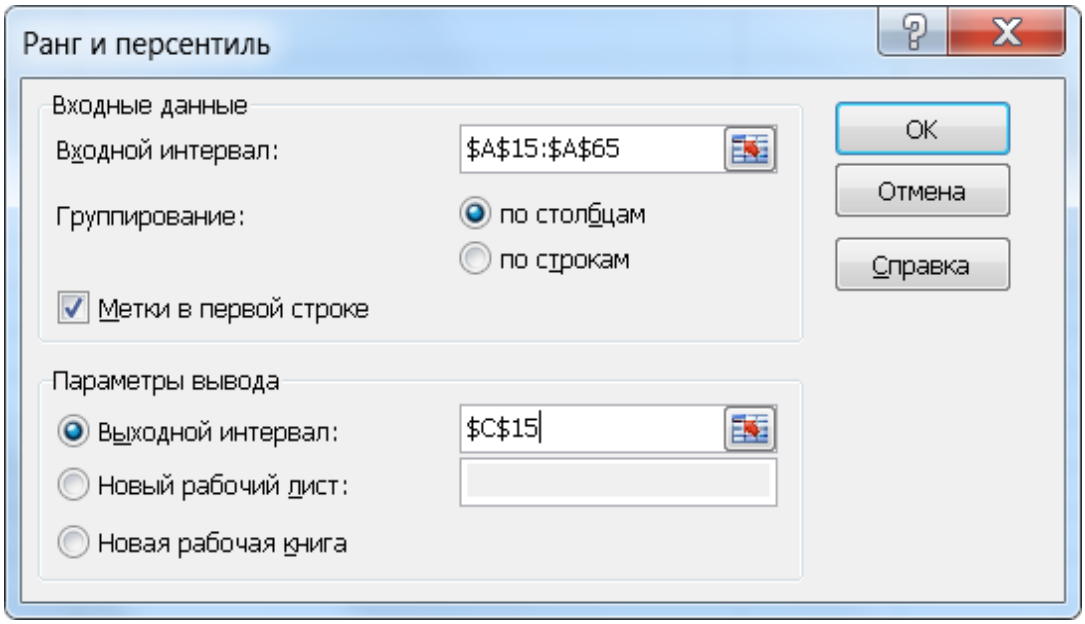


Рис. 22. Вид диалогового окна процедуры «Ранг и персентиль»

Пример применения процедуры

Дано: Зарплаты медсестер поликлиники**.**

Требуется: Провести количественный анализ относительного взаиморасположения данных в представленном наборе.

Для решения задачи используем процедуру «Ранг и персентиль». Значения параметров, установленных в одноименном диалоговом окне, приведены на рисунке 23, а исходные данные задачи и рассчитанные порядковые и процентные ранги для каждого значения из набора данных — на рисунке 24 (графы *Ранг* и *Процент).*

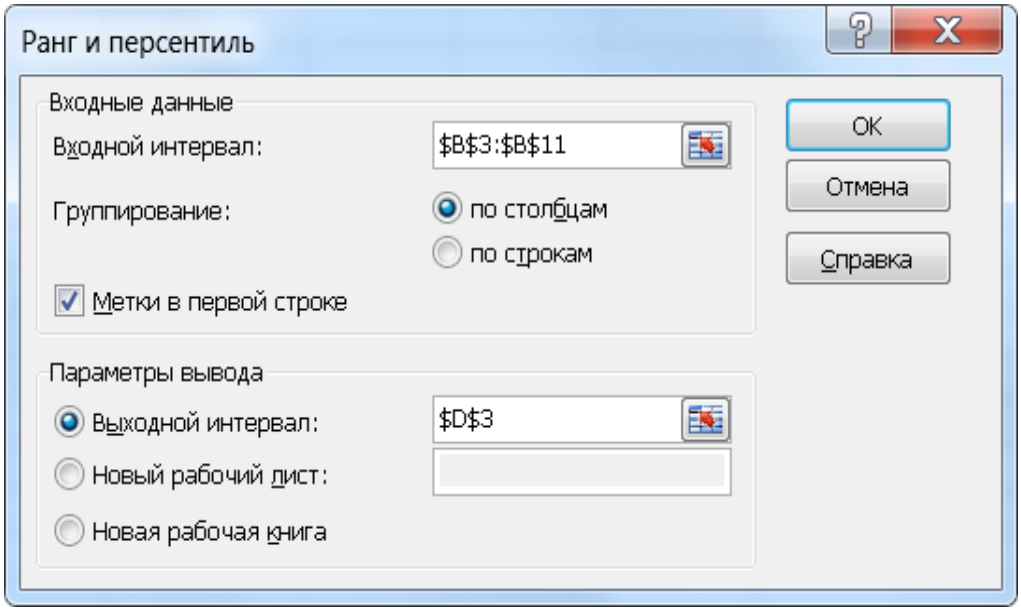


Рис. 23. Диалоговое окно процедуры «Ранг и персентиль»

3	Медсестры		*Точка*	*Медсестры*	*Ранг*	*Процент*
4	2100		1	2100	1	85,70%
5	2100		2	2100	1	85,70%
6	2000		3	2000	3	42,80%
7	2000		4	2000	3	42,80%
8	2000		5	2000	3	42,80%
9	1900		6	1900	6	28,50%
10	1800		7	1800	7	0,00%
11	1800		8	1800	7	0,00%

Рис. 24. Исходные данные и результаты работы процедуры «Ранг и персентиль»

✓ **Обратите внимание**! Максимальным значениям присвоен минимальный ранг, равным значениям присвоен одинаковый ранг. Последнее учитывается при определении рангов следующих чисел.

3.4. ФОРМИРОВАНИЕ РАНДОМИЗИРОВАННОЙ ВЫБОРКИ И ИЗУЧЕНИЕ ЕЕ СВОЙСТВ С ПОМОЩЬЮ ИНСТРУМЕНТОВ ПАКЕТА АНАЛИЗ ДАННЫХ

ЦЕЛИ ЗАНЯТИЯ

1. Ознакомиться с основными процедурами пакета Анализ данных, позволяющими формировать репрезентативную выборку и получать анализ распределения ее вариант.
2. Научиться использовать процедуру «Генерация случайных чисел» для формирования числовых последовательностей, подчиняющихся различным законам частотных распределений.
3. Изучить возможности формирования выборки из массивов данных с помощью инструмента «Выборка».
4. Осознать возможности применения процедуры «Ранг и персентиль» для анализа распределения вариант в исследуемой выборке.
5. Сформировать навыки использования инструмента «Гистограмма» для построения интервальных гистограмм частотного распределения выборки.

МЕТОДИКА ВЫПОЛНЕНИЯ РАБОТЫ

Постановка задачи 1

Дано: Номера историй болезни пациентов с определенной патологией в отделении больницы за текущий квартал находятся в диапазоне 300 — 950.

Требуется:

Выбрать случайным образом номера 40 историй болезни для формирования 3-х рандомизированных выборок. Из номеров историй болезни сформировать три массива данных, используя процедуру «***Генерация случайных чисел»*** со стартовыми значениями 1, 2, 3.

Постановка задачи 2

Дано: Для проведения серии лабораторных анализов имеется 10 образцов с номерами от 1 до 10.

Требуется:

1. Составить план эксперимента на месяц (30 дней) так, чтобы образцы чередовались в случайном порядке. Используйте процедуру «***Генерация случайных чисел»*** *(Распределение дискретное).*

2. Проверьте, сколько раз в месяц, встречается каждый образец, используя функцию *СЧЕТЕСЛИ().*

Постановка задачи 3

Дано: Характеристики выборки с нормальным частотным распределением. Среднее = 1 и стандартное отклонение = 1,5, число случайных чисел 30.

Требуется:

1. Для визуализации частотного распределения сгенерировать выборку с *нормальным* частотным распределением и заданными параметрами, используя процедуру «***Генерация случайных чисел»*** *(Распределение нормальное).*

2. По данным, полученным в п.1, построить с помощью процедуры «***Гистограмма»*** гистограмму частотных распределений.

Постановка задачи 4

Дано: LgA — результаты анализов военнослужащих, характеризующие состояние иммунной системы. (Lg A — иммуноглобулин А [повышается после заболевания и держится до года] — инфекция была давно).

Требуется:

1. Сформировать из предложенной совокупности *случайным* образом выборку из 30 значений, используя процедуру «***Выборка».***

2. Сформировать 3 выборки, применив к данным из предложенной совокупности процедуру ***«Выборка»****, метод выборки периодический* с шагом: 2, 3, 4.

Постановка задачи 5

Дано: Массивы, сформированные в результате решения задачи 4.

Требуется:

1. По результатам выполнения задачи 4 для сформированных выборок получить описательные статистики.

2. Проанализировав полученные характеристики, сделать вывод о нормальности частотных распределений и о том, какие методы в дальнейшем следует применять для статистического анализа этих выборок.

3. Вычислить доверительные интервалы для каждой из выборок. Проанализировав наличие перекрытия доверительных интервалов, сделать вывод: принадлежат ли эти выборки к одной генеральной совокупности или к различным (предположив, что частотные распределения во всех выборках нормальные).

Постановка задачи 6

Дано: Массивы, сформированные в задаче 4.

Требуется:

1. В выборках, сформированных в задаче 4, ранжировать данные с помощью процедуры «***Ранг и персентиль***».
2. Построить точечные гистограммы распределения данных.
3. Сравнить полученные диаграммы и определить, в какой из выборок меньше размах значений. Как влияет размер выборки на ее соответствие генеральной совокупности (количественную репрезентативность)?

Постановка задачи 7

Дано: Выборки, сформированные в результате решения задачи 4.

Требуется:

1. По результатам выполнения задачи 4 для сформированных выборок построить гистограммы частотных распределений.
2. Сравнить полученные результаты и определить, в какой из выборок данные находятся в меньшем диапазоне значений?

ХОД РАБОТЫ

Изучение процедур формирования и ранжирования рандомизированной выборки

1. Скопируйте в свою папку из папки ***Z: \Материалы для работы\Статистика*** книгу Microsoft Excel с именем *Пр.зан.№3 — Выборка.xls.*
2. Переименуйте скопированный файл, задав в качестве имени файла номер практической работы и свою фамилию, номер группы. Например, *Пр.зан.№3 Иванов А. — 24 леч.*
3. Изучите процедуры, приведенные в таблице 5, используя справочную систему программы, запишите в тетрадь в указанную таблицу найденную информацию по каждой процедуре.

Для этого примените команду: ***Справка ⇒ Анализ возможных вариантов ⇒ Выполнение статистического и инженерного анализа с помощью надстройки «Пакет анализа».***

Таблица 5.

Назначение процедур пакета Анализ данных

Название инструмента (процедуры)	Назначение процедуры
Генерация случайных чисел	
Выборка	
Ранг и персентиль	
Описательная статистика	

Применение процедуры «Генерация случайных чисел» для формирования рандомизированной выборки

Задача 1

Дано: Истории болезни (300-950). Вид распределения — равномерное.

Требуется:

Выбрать случайным образом **40** номеров историй болезни пациентов в диапазоне **300 — 950** для формирования 3-х рандомизированных выборок, используя процедуру «***Генерация случайных чисел»*** со стартовыми значениями 1, 2, 3.

Решение задачи 1

1. Перейдите на лист с именем *«Генер.сл.ч. 1».*
2. Вызовите процедуру «***Генерация случайных чисел».***
3. Задайте параметры процедуры в соответствии с рисунком 25. Результат разместите с адреса **A5**.

Учитывая возможность наличия повторов, *число случайных чисел* можно задать немного больше (44).

4. Повторите процедуру со стартовыми числами:
   - *случайное рассеивание* **2** — массив разместите с адреса **C5,**
   - *случайное рассеивание* **3** — массив разместите с адреса **E5.**

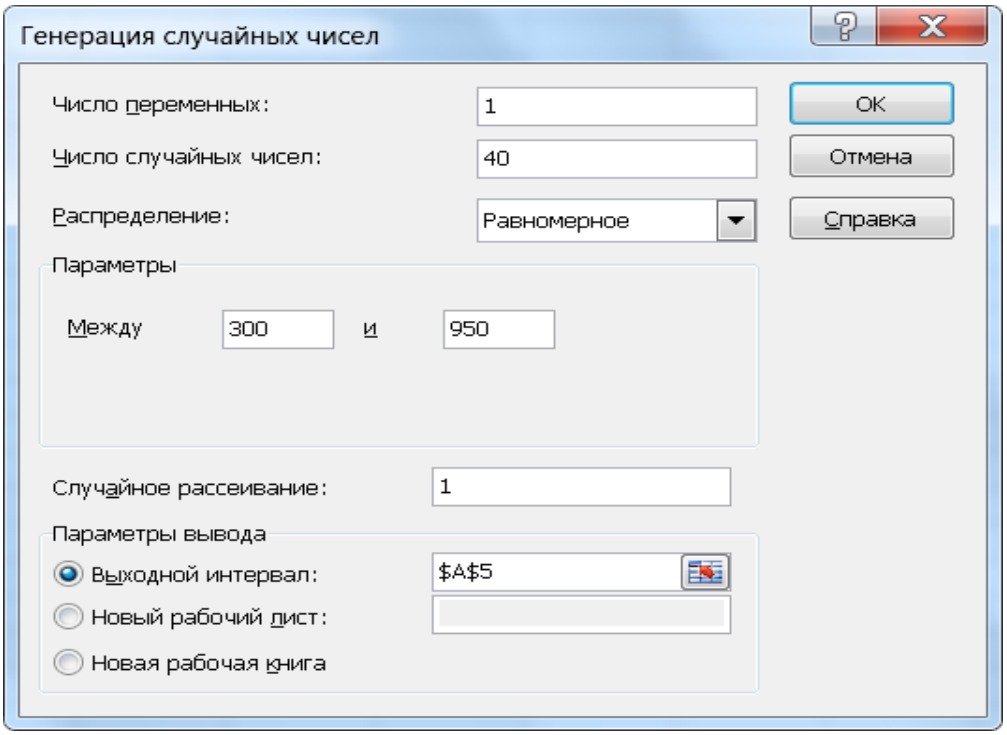
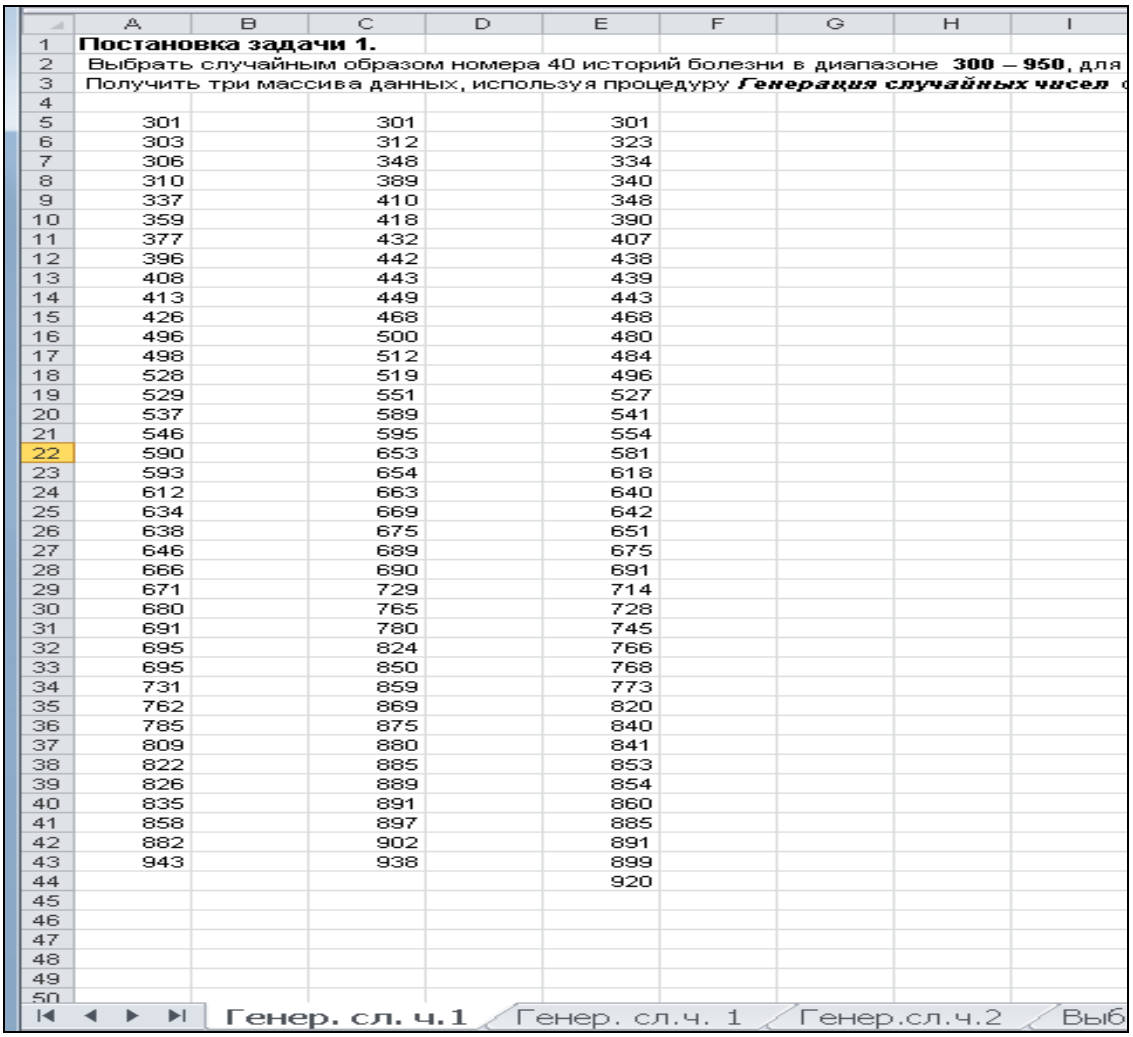


Рис. 25. Окно параметров процедуры «*Генерация случайных чисел»*

5. Округлите полученные значения в массивах до целых.
6. Упорядочьте полученные значения в каждом массиве по возрастанию.

7. Удалите повторы. Вид полученных результатов представлен на рисунке 26.



	A	B	C	D	E	F	G	H	I
1	**Постановка задачи 1.**								
2	Выбрать случайным образом номера 40 историй болезни в диапазоне **300 – 950**, для								
3	Получить три массива данных, используя процедуру ***Генерация случайных чисел***								
4									
5	301		301		301				
6	303		312		323				
7	306		348		334				
8	310		389		340				
9	337		410		348				
10	359		418		390				
11	377		432		407				
12	396		442		438				
13	408		443		439				
14	413		449		443				
15	426		468		468				
16	496		500		480				
17	498		512		484				
18	528		519		496				
19	529		551		527				
20	537		589		541				
21	546		595		554				
22	590		653		581				
23	593		654		618				
24	612		663		640				
25	634		669		642				
26	638		675		651				
27	646		689		675				
28	666		690		691				
29	671		729		714				
30	680		765		728				
31	691		780		745				
32	695		824		766				
33	695		850		768				
34	731		859		773				
35	762		869		820				
36	785		875		840				
37	809		880		841				
38	822		885		853				
39	826		889		854				
40	835		891		860				
41	858		897		885				
42	882		902		891				
43	943		938		899				
44					920				
45									
46									
47									
48									
49									
50									

Генер. сл. ч.1 | Генер. сл.ч. 1 | Генер.сл.ч.2 | Выб

Рис. 26. Вид листа с результатами решения задачи 1

8. Сравните полученные результаты.

9. Для проведения исследования воспользуйтесь историями болезни в соответствии с данными одного из полученных массивов.

Задача 2

Дано: Для проведения серии лабораторных анализов имеется 10 образцов с номерами от 1 до 10.

Требуется:

Составить план эксперимента на месяц (30 дней) так, чтобы образцы чередовались в случайном порядке, используя инструмент *«Генерация случайных чисел» (Распределение дискретное)*. Проверить, сколько раз в месяц встречается каждый образец, используя функцию *СЧЕТЕСЛИ()*.

Решение задачи 2

1. Перейдите на лист с именем *«Генер.сл.ч. 2»*. На листе представлены номера образцов и желаемые вероятности (P = 1/10).

2. Вызовите процедуру «Генерация случайных чисел» с распределением «дискретное», заполните поля окна процедуры в соответствии с рисунком 27.

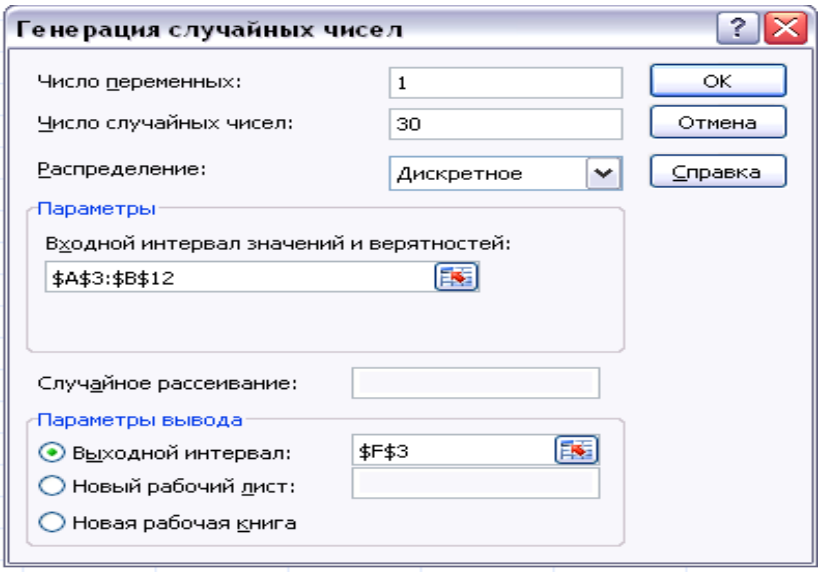


Рис. 27. Окно параметров процедуры «Генерация случайных чисел» с дискретным распределением

3. Вычислите сколько раз в течение месяца будет использоваться каждый образец. Для этого в ячейку ***C3*** введите формулу ***=СЧЕТЕСЛИ($F$3:$F$32;A3)***, выполните ее репликацию до ячейки ***C12*** включительно.

4. Вычислите итог, полученное значение должно быть равно 30.

	A	B	C	D	E	F
1		**План эксперимента**				
2	**Номер образца**	**Вероятнсть использования образца**	**проверка, сколько раз в месяц использовался образец**		**номер дня месяца**	**номер, используемого образца**
3	1	0,1	4		1	4
4	2	0,1	3		2	2
5	3	0,1	4		3	6
6	4	0,1	6		4	9
7	5	0,1	3		5	9
8	6	0,1	2		6	10
9	7	0,1	0		7	1
10	8	0,1	0		8	5
11	9	0,1	4		9	9
12	10	0,1	4		10	2
13		итого:	30		11	3
14					12	1
15					13	1
16					14	2
17					15	3
18					16	1
19					17	3
20					18	4
21					19	6
22					20	4
23					21	4
24					22	4
25					23	10
26					24	5
27					25	5
28					26	4
29					27	10
30					28	9
31					29	10
32					30	3

Генер.сл.ч. 1 | Генер.сл.ч. 2

Рис. 28. Результаты планирования эксперимента на месяц на Рабочем листе *«Генер.сл.ч. 2»*

5. Сделайте вывод, насколько равномерно распределены образцы по месяцу. Вид результатов представлен на рисунке 28.

Задача 3

Дано: Характеристики выборки с нормальным частотным распределением. Среднее = 1 и стандартное отклонение = 1,5, число случайных чисел 30.

Требуется:

1. По заданному среднему значению и стандартному отклонению сгенерировать массив данных с нормальным частотным распределением: ***Среднее = 1*** и ***Стандартное отклонение = 1,5***.
2. По полученным данным построить гистограмму частотного распределения.
3. Для сгенерированного массива данных вычислить с помощью статистических функций Microsoft Excel: среднее значение, стандартное отклонение, медиану, моду.

Решение задачи 3

1. Перейдите на лист с именем *«Генер.сл.ч. 3»*.
2. Вызовите процедуру «***Генерация случайных чисел***» с распределением «***нормальное***», укажите значения ***случайного рассеивания 2, среднее, стандартное отклонение, число случайных чисел 30.***
3. Для проверки результатов генерации вычислите с помощью соответствующих функций *среднее* и *стандартное отклонение, медиану и моду*.
4. Постройте интервальную гистограмму частотных распределений с помощью процедуры «***Гистограмма***», задайте самостоятельно ***значения карманов от -3 до +4 с шагом 1***. Вид результатов работы представлен на рисунке 29.

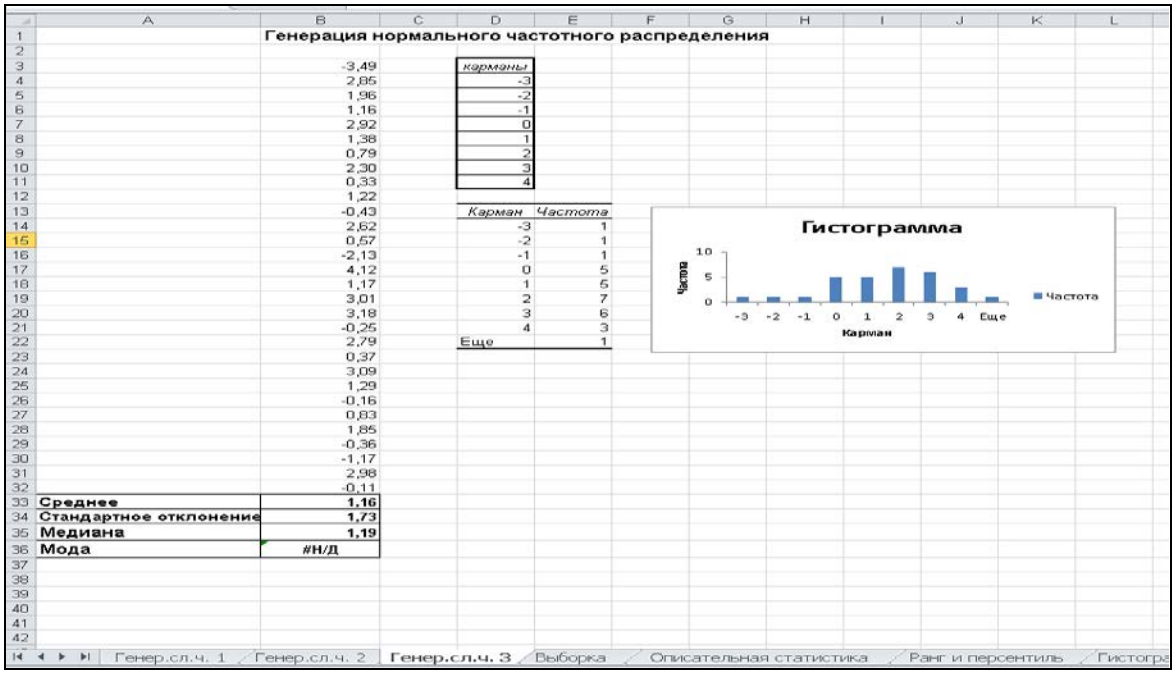


Рис. 29. Вид результатов решения задачи 3 с первой выборкой

5. Сформируйте еще два массива с другими значениями рассеивания. Выполните 3-й и 4-й пункты. Какие из них более всего соответствуют поставленной задаче?

Применение процедуры «Выборка» для формирования рандомизированной выборки

Задача 4

Дано: LgA — результаты обследования военнослужащих, характеризующие состояние иммунной системы. lg A — иммуноглобулин А (повышается после заболевания и держится до года) — инфекция была давно.

Требуется:

Из заданного массива сформировать:

1) случайным образом выборку из 30 значений, используя процедуру «***Выборка»;***

2) 3 выборки, применив к данным из предложенной совокупности процедуру ***«Выборка»****, метод выборки периодический* с шагом: 2, 3, 4.

Решение задачи 4

1. Перейдите на лист с именем «*Выборка*».
2. Вызовите процедуру ***«Выборка»,*** используя метод выбора *случайный,* сформируйте выборку из 30 значений, результаты разместите с ячейки ***E7***.
3. Вызовите процедуру ***«Выборка»***, используя метод выбора *периодический,* последовательно задавая шаги ***2, 3, 4***, разместите результаты соответственно с ячеек ***F7, G7, H7.*** Вид окна процедуры представлен на рисунке 30.

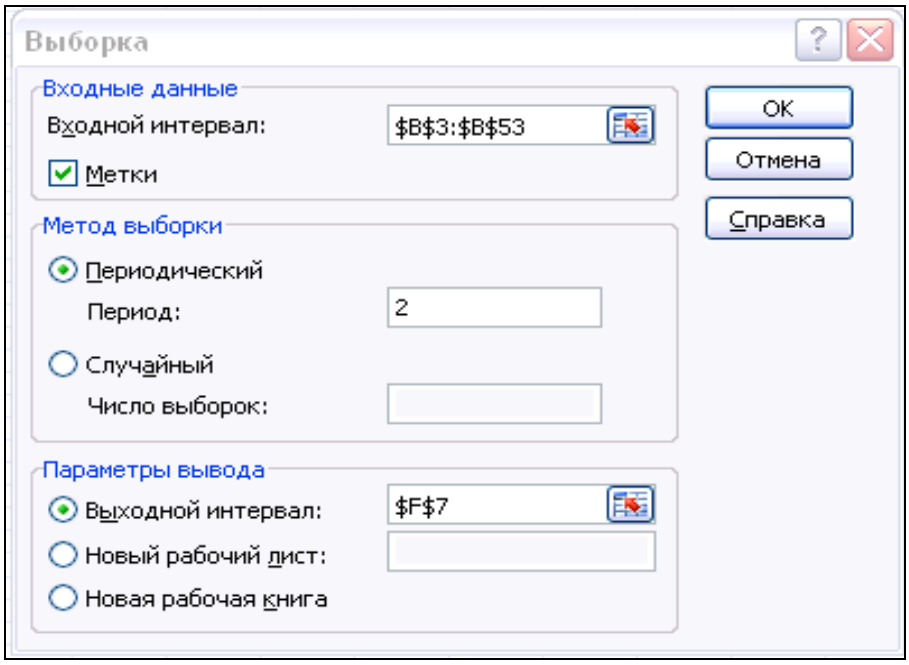
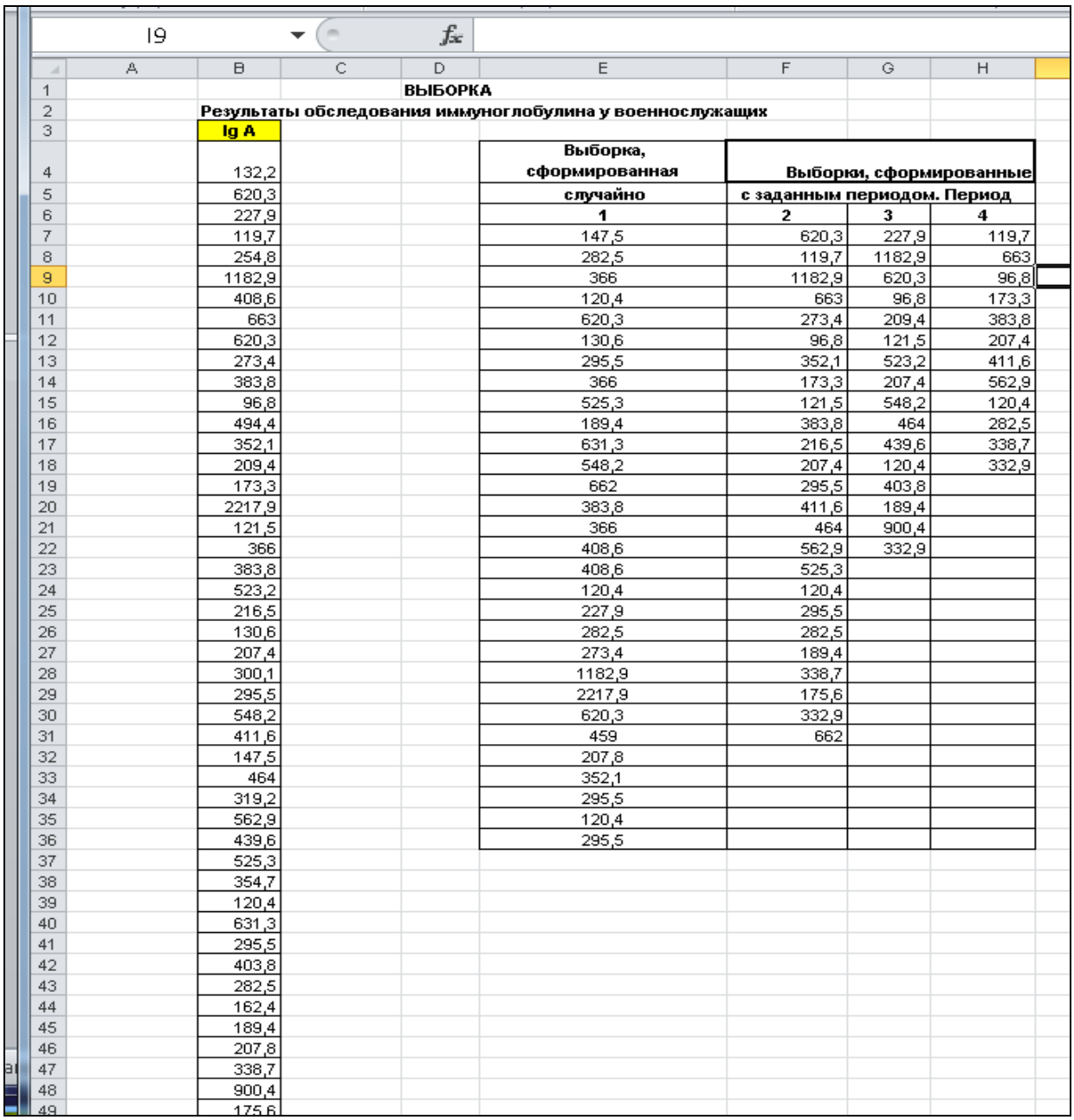


Рис. 30. Вид окна процедуры *«Выборка»*

4. Как заданный период влияет на объем полученной выборки.



	A	B	C	D	E	F	G	H
1				ВЫБОРКА				
2		Результаты обследования иммуноглобулина у военнослужащих						
3		lg A						
4		132,2			Выборка, сформированная	Выборки, сформированные		
5		620,3			случайно	с заданным периодом. Период		
6		227,9			1	2	3	4
7		119,7			147,5	620,3	227,9	119,7
8		254,8			282,5	119,7	1182,9	663
9		1182,9			366	1182,9	620,3	96,8
10		408,6			120,4	663	96,8	173,3
11		663			620,3	273,4	209,4	383,8
12		620,3			130,6	96,8	121,5	207,4
13		273,4			295,5	352,1	523,2	411,6
14		383,8			366	173,3	207,4	562,9
15		96,8			525,3	121,5	548,2	120,4
16		494,4			189,4	383,8	464	282,5
17		352,1			631,3	216,5	439,6	338,7
18		209,4			548,2	207,4	120,4	332,9
19		173,3			662	295,5	403,8	
20		2217,9			383,8	411,6	189,4	
21		121,5			366	464	900,4	
22		366			408,6	562,9	332,9	
23		383,8			408,6	525,3		
24		523,2			120,4	120,4		
25		216,5			227,9	295,5		
26		130,6			282,5	282,5		
27		207,4			273,4	189,4		
28		300,1			1182,9	338,7		
29		295,5			2217,9	175,6		
30		548,2			620,3	332,9		
31		411,6			459	662		
32		147,5			207,8			
33		464			352,1			
34		319,2			295,5			
35		562,9			120,4			
36		439,6			295,5			
37		525,3						
38		354,7						
39		120,4						
40		631,3						
41		295,5						
42		403,8						
43		282,5						
44		162,4						
45		189,4						
46		207,8						
47		338,7						
48		900,4						
49		175,6						

Рис. 31. Вид Рабочего листа *«Выборка»* с результатами решения задачи 4

Применение процедуры «Описательная статистика» для вычисления основных характеристик выборки

Задача 5

Дано: 4 массива, полученных в результате выполнения процедуры «Выборка» (см. задачу 4)

Требуется:

Для выборок, сформированных в результате выполнения задачи 4, вычислить описательные статистики, используя процедуру «***Описательная***

статистика», верхнюю и нижнюю границы доверительных интервалов.

Проанализировав полученные значения, сделать вывод о нормальности частотных распределений и методах, которые следует применять для последующего статистического анализа этих выборок.

Решение задачи 5

1. Перейдите на Рабочий лист с именем *«Описательная статистика»***.**

2. Вызовите процедуру **«*Описательная статистика»*** пакета Анализ данных. Вычислите описательные статистики для всех массивов данных, расположенных на активном Рабочем листе. Результаты разместите с ячейки ***C37*** в соответствии с рисунком 32.

3. Вычислите для каждого массива данных верхнюю и нижнюю границы доверительных интервалов.

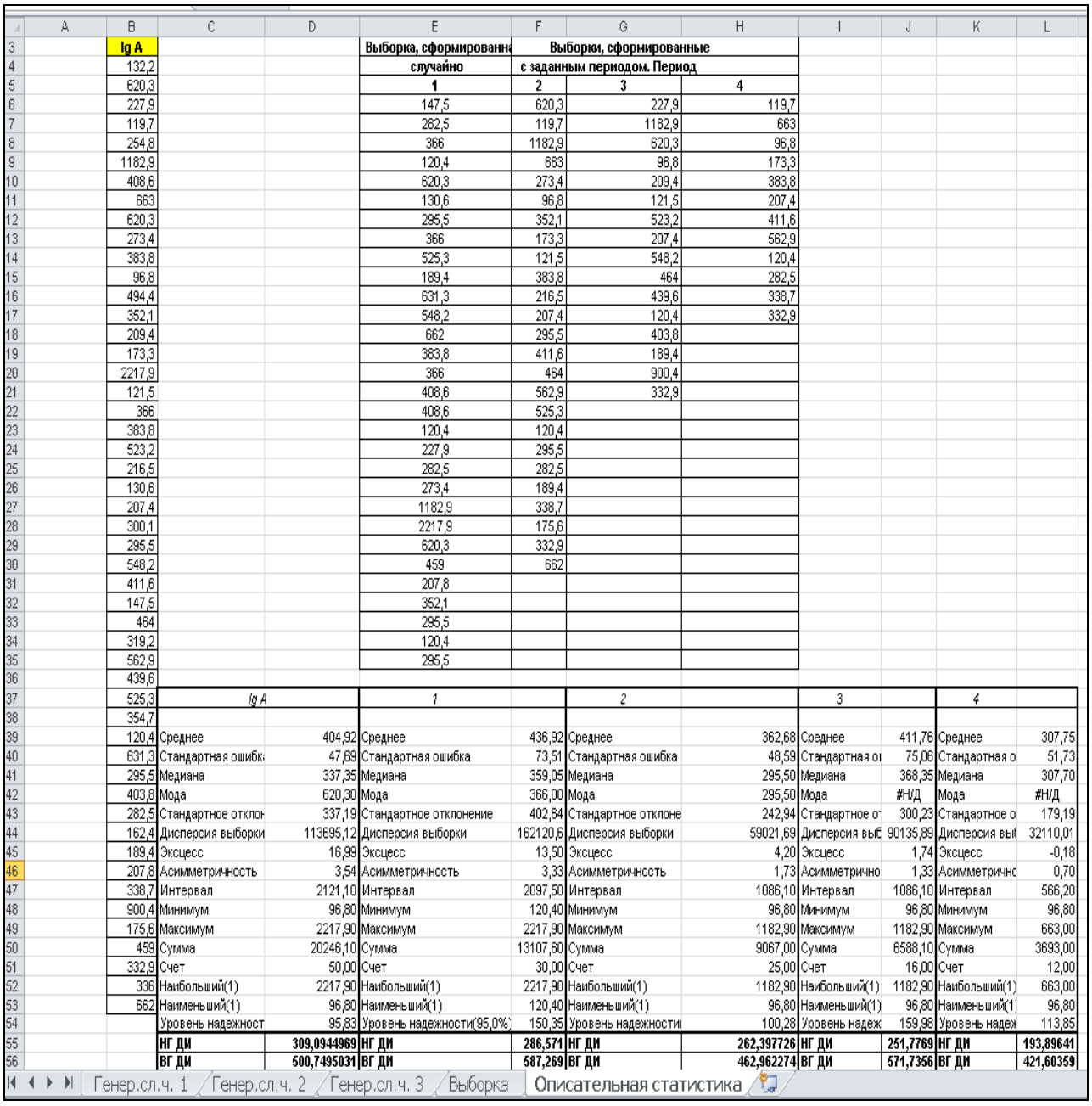
4. Проанализируйте полученные выборки на нормальность частотных распределений. Результаты запишите в тетрадь в таблицу 6.

Таблица 6.

Результаты анализа описательных статистик

Номер выборки	**Среднее**	**Медиана**	**Мода**	**Нормальность (да/нет)**	**Методы обработки (параметрические /непараметрические)**	**НГ ДИ**	**ВГ ДИ**	**Перекрываются ДИ (да/нет)**
LgA								
1								
2								
3								
4								

5. Проанализируйте, перекрываются ли доверительные интервалы полученных совокупностей. Сделайте вывод о возможности наличия общего среднего в этих выборках, предположив, что во всех выборках нормальное частотное распределение.



	lg A			Выборка, сформированна	Выборки, сформированные		
				случайно	с заданным периодом. Период		
				1	2	3	4
	132,2			147,5	620,3	227,9	119,7
	620,3			282,5	119,7	1182,9	663
	227,9			366	1182,9	620,3	96,8
	119,7			120,4	663	96,8	173,3
	254,8			620,3	273,4	209,4	383,8
	1182,9			130,6	96,8	121,5	207,4
	408,6			295,5	352,1	523,2	411,6
	663			366	173,3	207,4	562,9
	620,3			525,3	121,5	548,2	120,4
	273,4			189,4	383,8	464	282,5
	383,8			631,3	216,5	439,6	338,7
	96,8			548,2	207,4	120,4	332,9
	494,4			662	295,5	403,8	
	352,1			383,8	411,6	189,4	
	209,4			366	464	900,4	
	173,3			408,6	562,9	332,9	
	2217,9			408,6	525,3		
	121,5			120,4	120,4		
	366			227,9	295,5		
	383,8			282,5	282,5		
	523,2			273,4	189,4		
	216,5			1182,9	338,7		
	130,6			2217,9	175,6		
	207,4			620,3	332,9		
	300,1			459	662		
	295,5			207,8			
	548,2			352,1			
	411,6			295,5			
	147,5			120,4			
	464			295,5			
	319,2						
	562,9						
	439,6						

	lg A		1		2		3		4	
	525,3									
	354,7									
	120,4	Среднее	404,92	Среднее	436,92	Среднее	362,68	Среднее	411,76	Среднее	307,75
	631,3	Стандартная ошибк	47,69	Стандартная ошибка	73,51	Стандартная ошибка	48,59	Стандартная ої	75,06	Стандартная о	51,73
	295,5	Медиана	337,35	Медиана	359,05	Медиана	295,50	Медиана	368,35	Медиана	307,70
	403,8	Мода	620,30	Мода	366,00	Мода	295,50	Мода	#Н/Д	Мода	#Н/Д
	282,5	Стандартное отклон	337,19	Стандартное отклонение	402,64	Стандартное отклоне	242,94	Стандартное от	300,23	Стандартное о	179,19
	162,4	Дисперсия выборки	113695,12	Дисперсия выборки	162120,6	Дисперсия выборки	59021,69	Дисперсия выб	90135,89	Дисперсия выб	32110,01
	189,4	Эксцесс	16,99	Эксцесс	13,50	Эксцесс	4,20	Эксцесс	1,74	Эксцесс	-0,18
	207,8	Асимметричность	3,54	Асимметричность	3,33	Асимметричность	1,73	Асимметрично	1,33	Асимметричнс	0,70
	338,7	Интервал	2121,10	Интервал	2097,50	Интервал	1086,10	Интервал	1086,10	Интервал	566,20
	900,4	Минимум	96,80	Минимум	120,40	Минимум	96,80	Минимум	96,80	Минимум	96,80
	175,6	Максимум	2217,90	Максимум	2217,90	Максимум	1182,90	Максимум	1182,90	Максимум	663,00
	459	Сумма	20246,10	Сумма	13107,80	Сумма	9067,00	Сумма	6588,10	Сумма	3693,00
	332,9	Счет	50,00	Счет	30,00	Счет	25,00	Счет	16,00	Счет	12,00
	336	Наибольший(1)	2217,90	Наибольший(1)	2217,90	Наибольший(1)	1182,90	Наибольший(1)	1182,90	Наибольший(1)	663,00
	662	Наименьший(1)	96,80	Наименьший(1)	120,40	Наименьший(1)	96,80	Наименьший(1)	96,80	Наименьший(1	96,80
		Уровень надежност	95,83	Уровень надежности(95,0%)	150,35	Уровень надежности	100,28	Уровень надеж	159,98	Уровень надеж	113,85
		НГ ДИ	**309,0944969**	**НГ ДИ**	**286,571**	**НГ ДИ**	**262,397726**	**НГ ДИ**	**251,7769**	**НГ ДИ**	**193,89641**
		ВГ ДИ	**500,7495031**	**ВГ ДИ**	**587,269**	**ВГ ДИ**	**462,962274**	**ВГ ДИ**	**571,7356**	**ВГ ДИ**	**421,60359**

Генер.сл.ч. 1 | Генер.сл.ч. 2 | Генер.сл.ч. 3 | Выборка | Описательная статистика

Рис. 32. Вид Рабочего листа с результатами решения задачи 5

Применение процедуры «Ранг и персентиль» для вычисления рангов элементов выборки

Задача 6

Дано: Массивы, сформированные в задаче 4.

***Требуется*:**

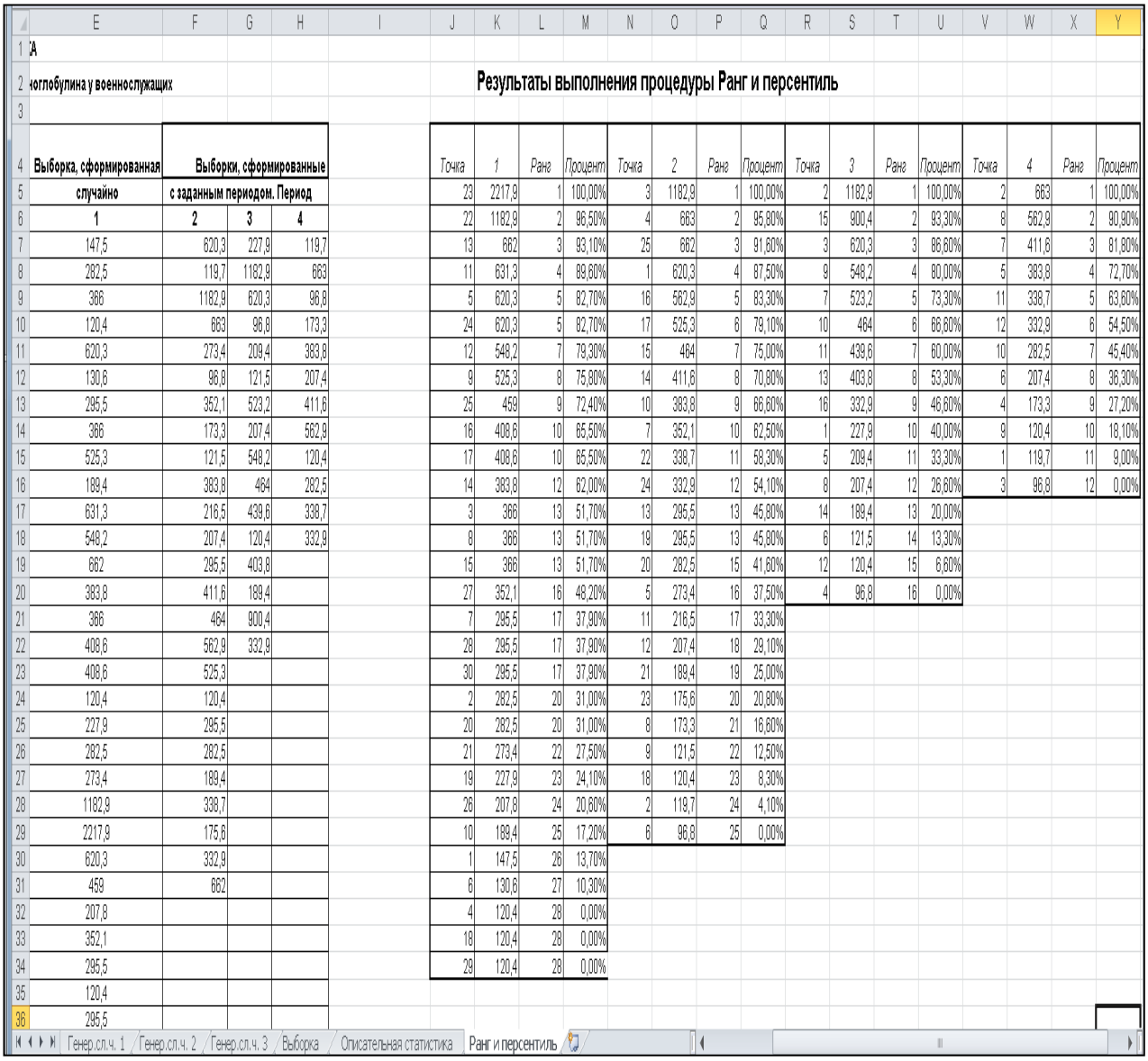
1. На Рабочем листе *«Ранг и персентиль»* в сформированных выборках ранжировать данные с помощью соответствующей процедуры.

2. Построить точечные гистограммы распределения данных. Сравнить полученные результаты и определить, в какой из выборок данные смещены в сторону больших значений.

Решение задачи 6

1. Перейдите на Рабочий лист *«Ранг и персентиль».*

2. Примените к данным выборок с номерами 1, 2, 3, 4 несколько раз процедуру «***Ранг и персентиль»,*** считая заголовком каждой выборки ее номер**.** Результаты разместите на Рабочем листе в соответствии с рисунком 33.



'A

ноглобулина у военнослужащих

Результаты выполнения процедуры Ранг и персентиль

Выборка, сформированная случайно	Выборки, сформированные с заданным периодом. Период		
1	2	3	4
147,5	620,3	227,9	119,7
282,5	119,7	1182,9	663
366	1182,9	620,3	96,8
120,4	663	96,8	173,3
620,3	273,4	209,4	383,8
130,6	96,8	121,5	207,4
295,5	352,1	523,2	411,6
366	173,3	207,4	562,9
525,3	121,5	548,2	120,4
189,4	383,8	464	282,5
631,3	216,5	439,6	338,7
548,2	207,4	120,4	332,9
662	295,5	403,8	
383,8	411,6	189,4	
366	464	900,4	
408,6	562,9	332,9	
408,6	525,3		
120,4	120,4		
227,9	295,5		
282,5	282,5		
273,4	189,4		
1182,9	338,7		
2217,9	175,6		
620,3	332,9		
459	662		
207,8			
352,1			
295,5			
120,4			
295,5			

Точка	1	Ранг	Процент
23	2217,9	1	100,00%
22	1182,9	2	96,50%
13	662	3	93,10%
11	631,3	4	89,60%
5	620,3	5	82,70%
24	620,3	5	82,70%
12	548,2	7	79,30%
9	525,3	8	75,80%
25	459	9	72,40%
16	408,6	10	65,50%
17	408,6	10	65,50%
14	383,8	12	62,00%
3	366	13	51,70%
8	366	13	51,70%
15	366	13	51,70%
27	352,1	16	48,20%
7	295,5	17	37,90%
28	295,5	17	37,90%
30	295,5	17	37,90%
2	282,5	20	31,00%
20	282,5	20	31,00%
21	273,4	22	27,50%
19	227,9	23	24,10%
26	207,8	24	20,60%
10	189,4	25	17,20%
1	147,5	26	13,70%
6	130,6	27	10,30%
4	120,4	28	0,00%
18	120,4	28	0,00%
29	120,4	28	0,00%

Точка	2	Ранг	Процент
3	1182,9	1	100,00%
4	663	2	95,80%
25	662	3	91,60%
1	620,3	4	87,50%
16	562,9	5	83,30%
17	525,3	6	79,10%
15	464	7	75,00%
14	411,6	8	70,80%
10	383,8	9	66,60%
7	352,1	10	62,50%
22	338,7	11	58,30%
24	332,9	12	54,10%
13	295,5	13	45,80%
19	295,5	13	45,80%
20	282,5	15	41,60%
5	273,4	16	37,50%
11	216,5	17	33,30%
12	207,4	18	29,10%
21	189,4	19	25,00%
23	175,6	20	20,80%
8	173,3	21	16,60%
9	121,5	22	12,50%
18	120,4	23	8,30%
2	119,7	24	4,10%
6	96,8	25	0,00%

Точка	3	Ранг	Процент
2	1182,9	1	100,00%
15	900,4	2	93,30%
3	620,3	3	86,60%
9	548,2	4	80,00%
7	523,2	5	73,30%
10	464	6	66,60%
11	439,6	7	60,00%
13	403,8	8	53,30%
16	332,9	9	46,60%
1	227,9	10	40,00%
5	209,4	11	33,30%
8	207,4	12	26,60%
14	189,4	13	20,00%
6	121,5	14	13,30%
12	120,4	15	6,60%
4	96,8	16	0,00%

Точка	4	Ранг	Процент
2	663	1	100,00%
8	562,9	2	90,90%
7	411,6	3	81,80%
5	383,8	4	72,70%
11	338,7	5	63,60%
12	332,9	6	54,50%
10	282,5	7	45,40%
6	207,4	8	36,30%
4	173,3	9	27,20%
9	120,4	10	18,10%
1	119,7	11	9,00%
3	96,8	12	0,00%

Генер.сл.ч. 1 / Генер.сл.ч. 2 / Генер.сл.ч. 3 / Выборка / Описательная статистика / Ранг и персентиль

Рис. 33. Результаты выполнения процедуры «Ранг и персентиль»

3. Упорядочьте результаты работы процедуры «Ранг и персентиль» по каждой выборке по убыванию поля «Ранг».

4. По полям «%» и «номер выборки» (значение элемента) постройте точечные диаграммы для каждой выборки. Общий вид Рабочего листа с диаграммами представлен на рисунке 34.

5. Сравните полученные диаграммы и определите, в какой из выборок меньше размах значений. Как влияет размер выборки на ее соответствие генеральной совокупности (количественную репрезентативность)?

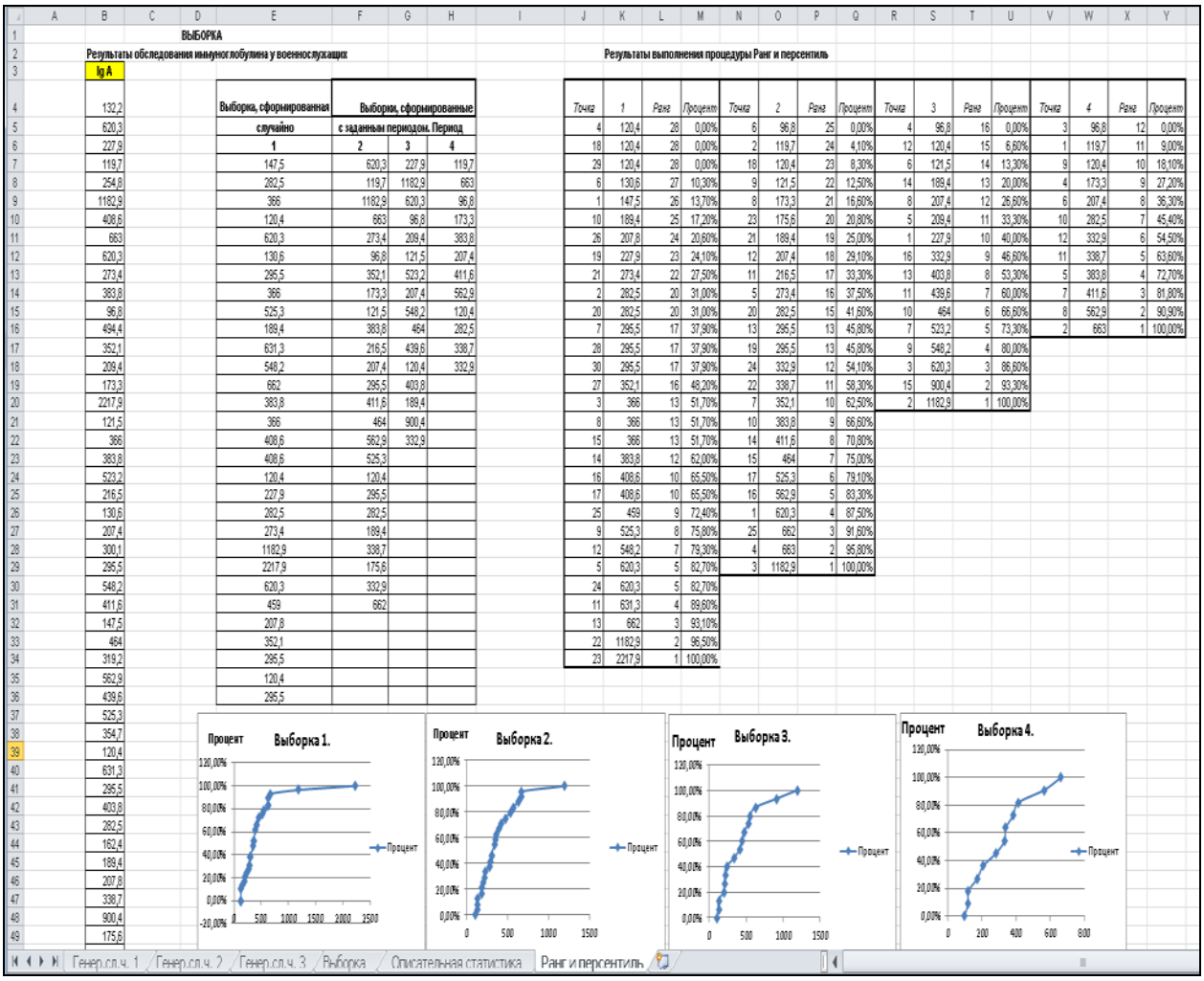


Рис. 34. Вид Рабочего листа *«Ранг и персентиль»* с точечными диаграммами

Применение процедуры «Гистограмма» для построения интервальных гистограмм частотного распределения элементов выборки

Задача 7

Дано: Выборки, сформированные в результате решения задачи 4.

***Требуется*:**

Для сформированных выборок построить гистограммы частотных распределений. Сравнить полученные результаты и определить, в какой из выборок данные находятся в меньшем диапазоне значений?

Решение задачи 7.

1. Перейдите на Рабочий лист *«Гистограмма».*

2. Для выборок с номерами 1, 2, 3, 4 постройте гистограммы частотных распределений с помощью процедуры ***«Гистограмма»,*** используя интервал карманов, представленный на Рабочем листе. Пример заполнения параметров диалогового окна для одной из выборок представлен на рисунке 35.

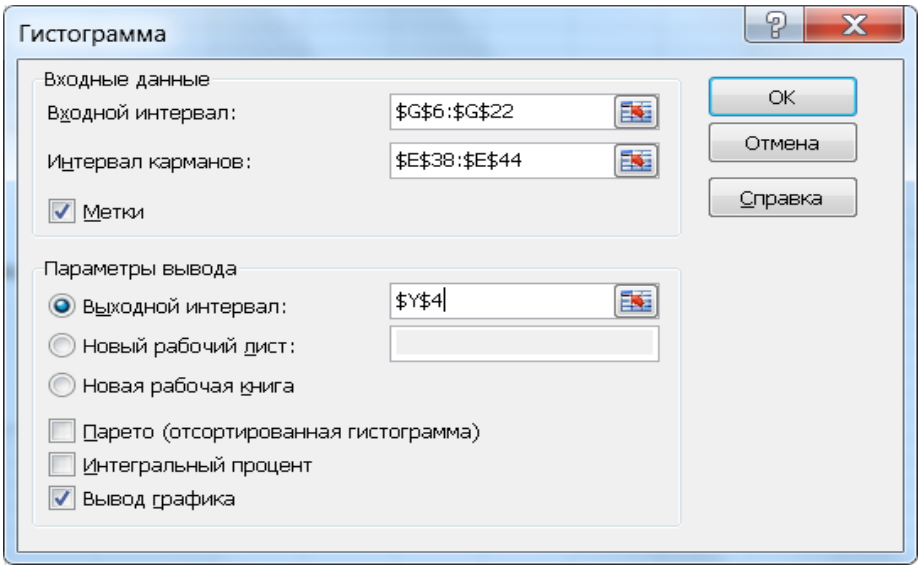


Рис. 35. Пример заполнения параметров диалогового окна «Гистограмма»

Вид результатов работы на Рабочем листе представлен на рисунке 36.

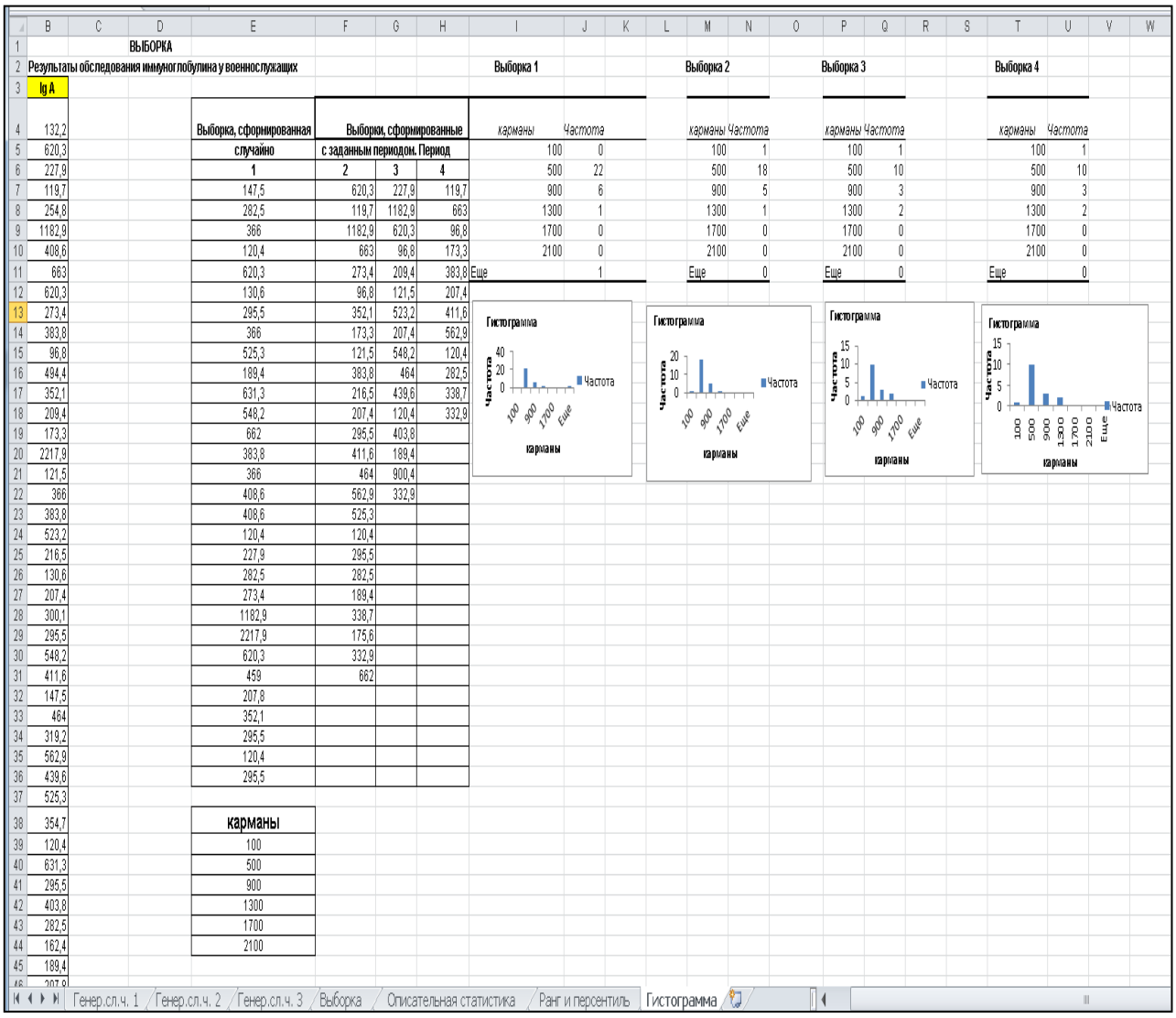


Рис. 36. Общий вид Рабочего листа *«Гистограмма»* с результатами работы

Вопросы для самоконтроля

1. Какие методы статистического анализа реализованы в пакете Анализ данных?

2. Назовите основные инструменты (процедуры) пакета Анализ данных, позволяющие формировать рандомизированную выборку.

3. Как обеспечить повторную генерацию выборки идентичной сформированной ранее с помощью процедуры «Генерация случайных чисел»?

4. Какие процедуры пакета Анализ данных позволяют изучить свойства выборки?

5. В чем различие процедур «Генерация случайных чисел» и «Выборка»?

6. Какие действия выполняет процедура «Гистограмма» пакета Анализ данных, кроме построения гистограммы частотного распределения выборки?

7. Перечислите основные процедуры пакета Анализ данных, реализующие критерий Стьюдента. Особенности применения этих процедур.

8. Назовите процедуры пакета Анализ данных, реализующие непараметрический критерий выявления достоверности различий χ^2 Пирсона.

9. Какая процедура пакета Анализ данных применяется для вычисления параметрического коэффициента корреляции Пирсона?

10. Какая процедура пакета Анализ данных может быть использована для вычисления непараметрического коэффициента ранговой корреляции Спирмена?

11. В чем заключается основной недостаток пакета Анализ данных?

12. Как в окнах процедур Анализ данных отражается способ выделения исходного массива: с заголовком, без заголовка?

13. Как в окнах процедур Анализ данных отражается способ размещения (по строкам или столбцам) исходного массива данных?

14. Как в окнах процедур Анализ данных указывается интервал ячеек, в которых отражаются результаты работы соответствующей процедуры?

4. МЕТОДЫ ВЫЯВЛЕНИЯ ДОСТОВЕРНОСТИ РАЗЛИЧИЙ

4.1. ТЕОРЕТИЧЕСКИЕ ОСНОВЫ

При выявлении достоверности различий используют понятия статистических гипотез, уровня значимости.

Понятие статистической гипотезы, уровня значимости, доверительного интервала

Под статистической гипотезой понимают всякое высказывание о генеральной совокупности (случайной величине), проверяемое по выборке (результатам наблюдений).

Проверяемую статистическую гипотезу принято называть **основной** (***нулевой*** *— 0 отличий*) гипотезой (обозначается $\mathbf{H_0}$), а противоречащую ей гипотезу — **альтернативной** *(или* ***конкурирующей)*** гипотезой (обозначается $\mathbf{H_1}$).

Нулевая гипотеза обычно утверждает, что между анализируемыми совокупностями (частотными распределениями) нет (0) статистически значимых различий.

По прикладному содержанию различают следующие основные виды высказываемых в ходе статистической обработки данных гипотез*:*

- *параметрические*: о числовых значениях исследуемой генеральной совокупности (среднее, медиана, и др.);
- *непараметрические:* о типе закона распределения исследуемой величины;
- *другие:*
  - об однородности двух или нескольких обрабатываемых выборок или некоторых характеристик анализируемых совокупностей;
  - о типе зависимости между компонентами исследуемого многомерного признака;
  - о независимости и стационарности обрабатываемого ряда наблюдений.

Если известна вероятность справедливости нулевой гипотезы (P_{H_0}), то выводы о ее принятии делают при сравнении этой вероятности с некоторой величиной, называемой *уровнем значимости*.

Уровень значимости **— максимальное значение** вероятности появления события, при котором событие (H_0) считается еще практически невозможным. В медицине наибольшее распространение получил уровень значимости (α), равный 0,05. Поэтому, если вероятность P_{H_0} (p), с которой интересующее событие может произойти случайным образом, $p < 0,05$, то принято считать это событие маловероятным, и если оно все же произошло, то это не было случайным.

Иначе говоря, чем *выше* уровень значимости, тем *меньше* можно доверять утверждению, что статистически значимые различия существуют.

Уровень значимости представляет собой вероятность ошибки, связанной с распространением наблюдаемого результата на всю генеральную совокупность, которую совершают, отвергая нулевую гипотезу, принимая альтернативную.

Верхняя граница $\alpha < 0,05$ статистической значимости содержит довольно большую вероятность ошибки (5%). Поэтому в тех случаях, когда требуется особая уверенность в достоверности полученных результатов, принимается уровень значимости $\alpha < 0,01$ или даже $\alpha < 0,001$ (токсикология, фармакология*)*.

В статистике используют понятия *доверительная вероятность* $P=1-\alpha$, это вероятность достаточная для того, чтобы с уверенностью судить о принятом статистическом решении.

При: $\alpha = 0,05$ $P = 95\%$;
$\alpha = 0,01$ $P = 99\%$;
$\alpha = 0,001$ $P = 99,9\%$.

Интервал, в котором с заданной доверительной вероятностью $P=1-\alpha$ находится оцениваемый параметр, называется ***доверительным интервалом (ДИ).***

Следует помнить, что если объем выборки небольшой, то частотное распределение не следует точно нормальному закону.

Выявление достоверности различий

Для выявления достоверности различий между выборками пользуются различными математическими критериями. Критерии выявления достоверности различий, как и гипотезы, могут быть параметрическими и непараметрическими.

К **параметрическим** критериям относится **критерий Стьюдента**, к непараметрическим — **критерий согласия Пирсона χ^2 (**Хи-квадрат).

Параметрические методы выявления достоверности различий

Параметрические критерии используются только тогда, когда обе выборки имеют нормальные законы частотных распределений.

В этом случае достоверность различия можно выявить, применив **критерий Стьюдента,** или осуществив анализ **доверительных интервалов средних** двух выборок.

Критерий Стьюдента (t) — наиболее часто используется для проверки гипотезы H_0: «*Средние двух выборок относятся к одной и той же генеральной совокупности*». Критерий позволяет найти **вероятность** этой гипотезы. Если эта вероятность $\boldsymbol{P_{H_0}}$ (p) ниже уровня значимости ($p < 0,05$), то принято считать, что выборки относятся к двум разным генеральным совокупностям, т.е. статистически значимо различаются.

Итак, для оценки достоверности события по критерию Стьюдента принимается нулевая гипотеза, что ***средние выборок статистически значимо не различаются.***

Функция ТТЕСТ()

В среде электронных таблиц Microsoft Excel критерий Стьюдента реализуется функцией **ТТЕСТ**(*М1;М2;Хв;Тип*). В более поздних версиях программы Microsoft Excel имя функции **СТЬЮДЕНТ.ТЕСТ**(*М1;М2;Хв;Тип*),

где **М1** — массив первой выборки (например, контрольная группа),

М2 — массив второй выборки (например, исследуемая группа),

Хв — хвосты распределения (может принимать значения 1 или 2),

Тип — тип теста (значения от 1 до 3).

Параметр ***Хв (хвосты)*** предусматривает возможность ввода значений: **1 —** одностороннее распределение**, 2 —** двустороннее**.**

Параметр ***Тип*** определяет вид выполняемого теста**:**

- **1 —** парный (используются пары данных**);**
- **2 —** двухвыборочный **(**сравнение двух разных выборок с равными дисперсиями);
- **3 —** двухвыборочный **(**разные выборки с неравными дисперсиями).

В тех случаях, когда используются две группы, состоящие из ***одних и тех же*** пациентов, применяется значение параметра Тип=**1,** ***разных*** **—** Тип=**3**.

При использовании критерия Стьюдента следует учитывать следующие ограничения на его применение:

1) этот критерий может применяться для работы как с малыми от 4 до 100 единиц, так и с большими выборками;

2) распределение элементов выборок должно подчиняться ***нормальному закону распределения*** *(Гаусса)*.

Пример использования функции ТТЕСТ( ) приведен на рисунке 37.

C9 =ТТЕСТ(C4:C8;D4:D8;2;3)

	A	B	C	D	E
1	**Температуры (С) двух групп больных**				
2	**№ задания**	**функция**	***номера групп***		
3			**1**	**2**	
4			37,3	37,1	
5			37,3	37,2	
6			37,4	37,3	
7			37,5	37,4	
8			37,6	37,5	
9		**ТТЕСТ**	0,228053		

Рис. 37. Вид фрагмента листа Excel с функцией ТТЕСТ( )

В качестве параметров функции ТТЕСТ(), представленной на рисунке 37, взяты интервалы размещения массивов группы 1 — *C4:C8*, группы 2 — *D4:D8*, значение параметра *Хвосты* равно *2* (двустороннее распределение), параметр *Тип* равен *3* (разные пациенты).

В результате применения критерия Стьюдента функция ТТЕСТ() возвращает (вычисляет) значение вероятности нулевой гипотезы $Рн_0$.

Алгоритм анализа полученного результата

Целью выполнения алгоритма анализа результатов работы функций, реализующих вычисление вероятности нулевой гипотезы (ТТЕСТ() и ХИ2ТЕСТ()), является выявление достоверности различий в двух выборках и определение на основании этого эффективности новой методики лечения (реабилитации) пациентов, эффективности нового фармацевтического препарата.

Приведем словесный и графический (рис.38) алгоритмы анализа данных, полученных с помощью функции ТТЕСТ().

Результатом выполнения предложенных ниже алгоритмов являются текстовые переменные Т1 и Т2. Текстовая переменная Т1 содержит информацию о достоверности различий двух выборок. Переменная Т2 — об эффективности новой методики (фармацевтического препарата).

Словесная форма алгоритма выявления эффективности новой методики

Начало алгоритма.

1. Применить функцию, ТТЕСТ() — параметрический критерий Стьюдента, выявляющий достоверность различий. (Функция вычисляет вероятность нулевой гипотезы ($Р_{Н_0}$));

2. Сравнить $Р_{Н_0}$ с уровнем значимости α= 0,05 (0,01; 0,001).

Если $Р_{Н_0} > \alpha$

то

справедлива нулевая гипотеза (H_0): данные в двух выборках (группах) достоверно не различаются, т.е. принадлежат одной и той же генеральной совокупности. Следовательно, новая методика лечения (фармацевтический препарат) неэффективна.

Т1:= «*Справедлива H_0, данные в двух выборках* ***достоверно не различаются***», Т2:= «*Новая методика неэффективна*».

Перейти к пункту 5.

иначе

($Р_{Н_0} \leq \alpha$) перейти к пункту 3.

3. *Справедлива альтернативная гипотеза, данные в двух группах* ***достоверно различаются***. Т1:= «*Справедлива H_1, две выборки достоверно различаются*».

4. Сравнить средние значения в двух выборках.

Если в ***исследуемой*** группе наблюдается ***улучшение*** признака

то

новая методика (фармацевтический препарат) **эффективна**,

Т2:= «*Новая методика эффективна*», перейти к пункту 5.

иначе

новая методика (фармацевтический препарат) неэффективна,

Т2:= «*Новая методика неэффективна*».

5. Вывести значения Т1 и Т2.

Конец алгоритма.

Графическая форма представления алгоритма на примере анализа результатов, применения критерия Стьюдента.

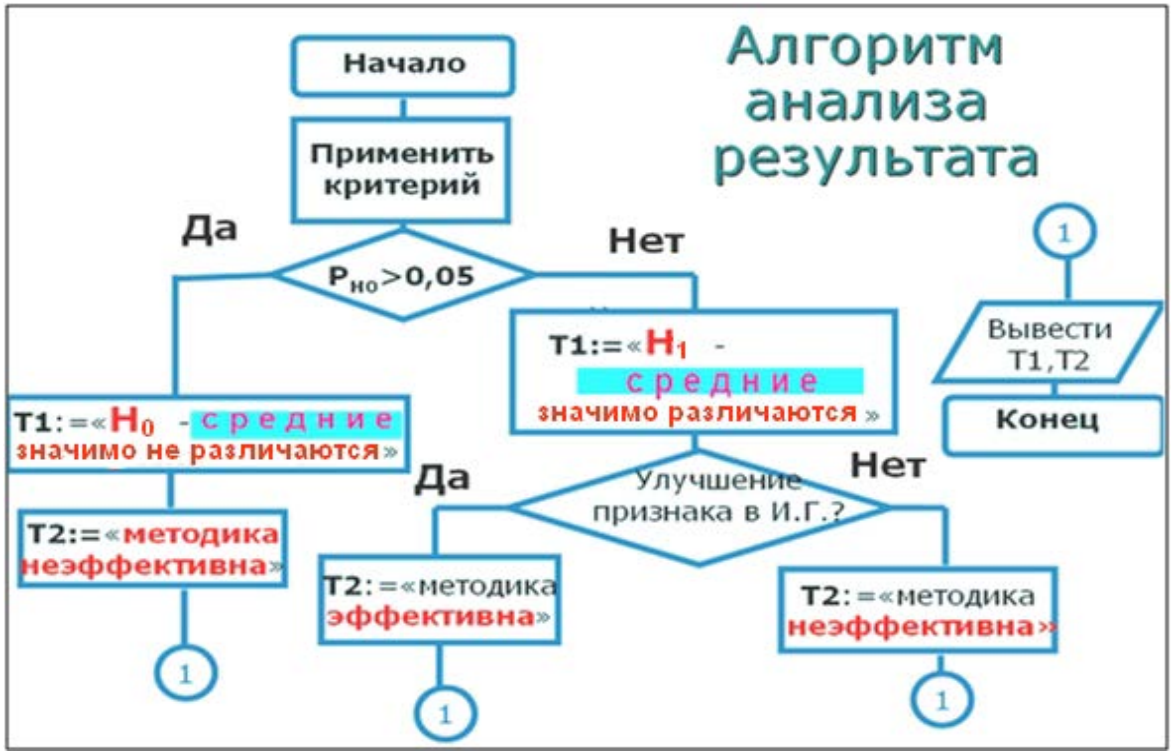


Рис. 38. Алгоритм анализа результатов применения функции ТТЕСТ() (ХИ2ТЕСТ())

Выводы по представленному на рисунке 37 значению.

Так как **$Рн_0$=0,228 > уровня значимости,** равного **0,05**, принимается нулевая гипотеза. Две группы статистически значимо не различаются (принадлежат к одной генеральной совокупности), следовательно, новая методика **неэффективна**.

Процедуры пакета Анализ данных, реализующие критерий Стьюдента

Аналогами функции ТТЕСТ() являются несколько процедур пакета Анализ данных, среди которых: *«Парный двухвыборочный тест для средних»,* «*Двухвыборочный t-тест с одинаковыми дисперсиями»*, *«Двухвыборочный t-тест с различными дисперсиями».*

Указанные процедуры реализуют те же вычисления, что и функция ТТЕСТ() при различных значениях параметра **Тип**.

«Парный двухвыборочный t-тест для средних» соответствует работе функции ТТЕСТ() при значении параметра **Тип=1**.

«***Двухвыборочный t-тест с одинаковыми дисперсиями»*** реализует алгоритм работы функции ТТЕСТ() при значении параметра **Тип=2**.

«Двухвыборочный t-тест с различными дисперсиями». реализует алгоритм работы функции ТТЕСТ() при значении параметра **Тип=3**.

Процедуры применимы только к выборкам с нормальным частотным распределением.

В диалоговых окнах этих процедур задаются параметры, представленные на рисунке 39

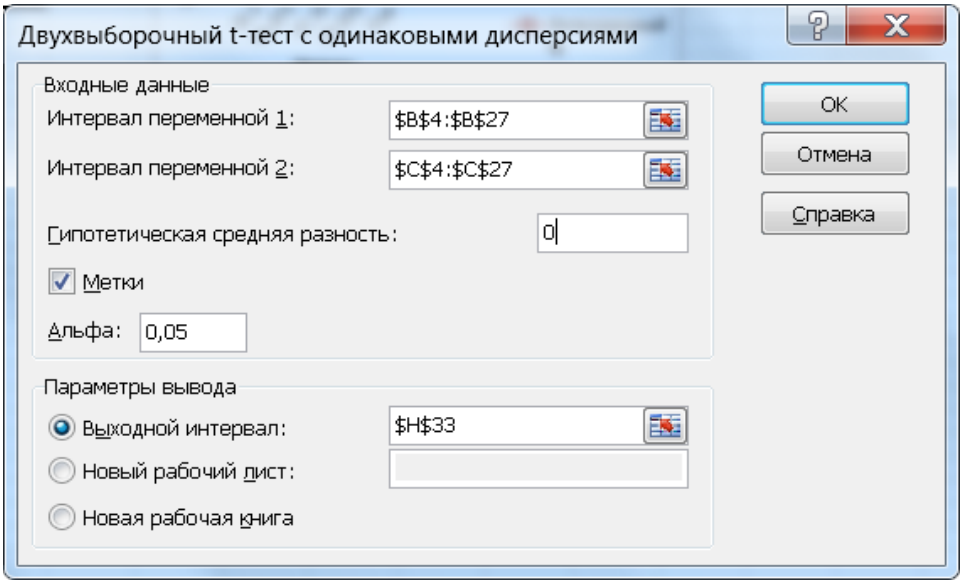


Рис. 39. Окно процедуры пакета Анализ данных, реализующей алгоритм функции ТТЕСТ()

Результаты работы функций будут иметь вид, близкий к представленному на рисунке 40.

Двухвыборочный t-тест с различными (одинаковыми) дисперсиями

	Контрольная	*Исследуемая*
Среднее	140,5	121,7
Дисперсия	184,2777778	144,0111111
Наблюдения	10	10
Гипотетическая разность средних	0	
df	18	
t-статистика	3,281177745	
P(T<=t) одностороннее	0,002075072	
t критическое одностороннее	1,734063607	
P(T<=t) двухстороннее	0,004150144	
t критическое двухстороннее	2,10092204	

Рис. 40. Пример результатов работы процедур пакета Анализ данных, реализующих критерий Стьюдента

Вероятность нулевой гипотезы представлена параметром **P(T<=t) двухстороннее**.

Сравнивая эту вероятность с уровнем значимости можно сделать вывод о том, какая из выдвинутых гипотез справедлива, а далее по рассмотренному алгоритму, можно сформулировать вывод об эффективности новой методики лечения.

Критерий выявления достоверности различий по доверительным интервалам для среднего в двух группах при нормальном частотном распределении

Если доверительные интервалы средних *двух выборок с нормальным частотным распределением* ***перекрываются,*** *то между выборками* ***нет статистически значимых различий.***

В случае, когда у доверительных интервалов двух выборок есть общая область, на ней может быть общее среднее, принадлежащее одной и той же генеральной совокупности.

Непараметрические методы выявления достоверности различий

Для определения достоверности различий существует множество непараметрических критериев. Одним из важнейших является критерий согласия Пирсона χ^2 (Хи-квадрат). Критерий согласия χ^2 используется, когда необходимо подтвердить или отвергнуть гипотезу о *совпадении законов частотных распределения* случайной величины в двух выборках.

✓ **Обратите внимание!** При применении параметрических критериев, в выборках используются значения исследуемого признака (температуры, давления, количества лейкоцитов, частоты сердечных сокращений и других).

При применении критерия Пирсона χ2 мы имеем дело с частотными распределениями, т.е. прежде тем приступить к применению этого критерия, надо от значений признака перейти к частотам его встречаемости на выбранных диапазонах значений, см. практическую работу №2, построение гистограммы.

При использовании критерия Пирсона χ^2 принимается нулевая гипотеза ($\mathbf{H_0}$) о том, что ***отсутствует достоверное различие в законах частотного распределения*** в двух выборках и альтернативная гипотеза ($\mathbf{H_1}$) о том, что эти ***законы статистически значимо различаются***.

Особенности применения критерия χ^2

1) Непараметрический критерий согласия Пирсона χ^2 получил большое распространение, так как дает возможность его использования с *различными формами распределений выборок*. Основное преимущество χ^2-критерия в его ***гибкости***. Этот критерий можно применять для проверки допущения о любом распределении, даже не зная параметров частотного распределения.

2) Основной недостаток этого критерия — *нечувствительность* к обнаружению адекватной модели, когда *число наблюдений невелико (меньше 5)*. В этом случае следует объединять (укрупнять) соседние интервалы значений признака до тех пор, пока частота не станет больше 3-5.

Критерий согласия Пирсона $\boldsymbol{\chi^2}$ в среде ЭТ Microsoft Excel реализует функция **ХИ2ТЕСТ(М1;М2),** в более новых версиях программы Microsoft

Excel — функция **ХИ2.ТЕСТ(М1;М2),**

где **М1 —** фактический интервал, **М2 —** ожидаемый интервал.

Эта функция вычисляет **вероятность нулевой гипотезы P_{H0}.**

Если полученная вероятность превосходит уровень значимости (0,05), то считают, что нулевая гипотеза справедлива, не противоречит опытным данным и может быть принята, т.е. частотные распределения в двух группах статистически значимо не различаются, следовательно, новая методика не эффективна. Рассмотрим пример применения функции ХИ2ТЕСТ(), реализующей критерий согласия Пирсона.

Постановка задачи

Дано: две группы пациентов: в контрольной группе **100** человек, в вакцинированной **— 50**. Во время эпидемии в *контрольной* группе заболело **60** человек, а в *вакцинированной* **— 5** человек.

***Требуется*:**

Определить эффективность действия вакцины.

Решение задачи

Нам известно *фактическое* распределение пациентов, для применения функции надо вычислить *теоретическое* **— *ожидаемое*** распределение.

При определении значений *ожидаемого распределения* вычисляют:

1) ***долю объектов с указанным значением,*** в данном случае заболевших в двух группах, как сумму заболевших, деленную на общее количество пациентов: (60+5)/(100+50) = 0,43 (1)

2) ***ожидаемое количество*** заболевших в каждой группе, как долю заболевших (1), умноженную на количество пациентов в конкретной группе:

- ожидаемое количество заболевших в **контрольной** группе равно 0,43*100=43 (человек);
- ожидаемое количество заболевших в **вакцинированной** группе равно 0,43*50=21,67 (человек);

3) зная количество заболевших и общее количество пациентов, можно определить ***количество здоровых*** в каждой группе:

- в контрольной группе количество здоровых равно:
  100 – 43 = 57 (человек)
- в вакцинированной — количество здоровых равно:
  50 – 21,67 = 28,33 (человека).

Понятно, что это математика, и количество людей в реальной жизни не может быть дробным.

Вид фрагмента листа ЭТ с решением этой задачи представлен на рисунке 41.

При вводе функции в качестве фактического интервала указывали ячейки **B5:C6**, в качестве ожидаемого — **B9:C10,**

=ХИ2ТЕСТ(B5:C6;B9:C10).

	A	B	C	D
1				
2	**Определение эффективности действия вакцины**			
3			**Количество больных**	
4	**фактическое**	**контрольная**	**вакцинированная**	
5	***больные***	**60**	**5**	
6	***здоровые***	**40**	**45**	
7				
8	**ожидаемые**			
9	***больные***	**43**	**21,67**	
10	***здоровые***	**57**	**28,33**	
11				
12	**доля больных=**	0,43		
13				
14				
15	**ХИ2ТЕСТ=**	**4,45E-09**		
16				

Рис. 41 Вид Рабочего листа с результатами применения функции ХИ2ТТЕСТ()

Результат вычислений указанной функции **4,46E-09** — вероятность нулевой гипотезы, следует интерпретировать, как $Рн_0$= **$4{,}45*10^{-09}$**.

Алгоритм анализа полученного результата

Результат выполнения функции ХИ2ТЕСТ() — непараметрического критерия, выявляющего достоверность различия частотных распределений в двух группах, **$РН_0= 4{,}45*10^{-09}$**.

Сравним **$РН_0$ с уровнем значимости α= 0,05** (0,01; 0,001).

Так как **$РН_0$ < α** справедлива **альтернативная гипотеза**, частотные распределения, а следовательно, данные в двух группах **достоверно различаются**.

Сравним ***относительное количество заболевших*** в двух выборках: в ***вакцинированной*** группе показатель относительное количество заболевших (5/50 = 0,1) ***лучше***, чем в ***контрольной*** (60/100 = 0,6), следовательно, ***вакцина эффективна***.

Следует отметить, что пакет Анализ данных включает только процедуры для параметрических методов, процедуры — аналога функции ХИ2ТЕСТ() в этом пакете нет.

4.2. ПАРАМЕТРИЧЕСКИЕ И НЕПАРАМЕТРИЧЕСКИЕ МЕТОДЫ ВЫЯВЛЕНИЯ ДОСТОВЕРНОСТИ РАЗЛИЧИЙ

ЦЕЛИ ЗАНЯТИЯ

1. Ознакомиться с основными понятиями выявления достоверности различий (основная и альтернативная гипотезы, уровень значимости, алгоритм выявления достоверности различий).
2. Получить навыки применения встроенной функции Microsoft Excel ТТЕСТ(), вычисляющей достоверность различий по критерию Стьюдента.
3. Научиться формулировать выводы о наличии достоверности различий в двух выборках, эффективности новой методики лечения или лекарственного средства.

4. Получить представление о возможности выявления достоверности различий по доверительным интервалам двух выборок.

5. Ознакомиться с алгоритмом применения непараметрического критерия Пирсона, реализуемого функцией ХИ2ТЕСТ().

МЕТОДИКА ВЫПОЛНЕНИЯ РАБОТЫ

Постановка задачи 1

Дано: Выборочные совокупности, содержащие сведения о температуре пациентов двух групп: контрольной и исследуемой.

Начало таблицы

Контрольная	***Исследуемая***
38,5	37,6
38,2	37,1
39	38,1
39,5	38,2
38,7	38
37	37,9
38,4	36,8
38,3	37,1
39,2	38,2
37,9	36,8
37,9	36,5
38,4	38,1

Продолжение таблицы

Контрольная	***Исследуемая***
38,6	38,2
38,4	36,7
37,3	36,9
37,7	36,8
37,4	37
37	37,5
38,5	37,6
37,9	37,2
37,8	37,3
38	37,6
38,8	38,2

Требуется:

1. Записать в тетрадь в таблицу 7 из встроенной справки программы Microsoft Excel назначение, синтаксис и примеры применения функций ТТЕСТ() и ХИ2ТЕСТ().

2. Вычислить описательные статистики и границы доверительных интервалов для каждой группы.

3. Выявить эффективность нового лекарственного средства, с помощью функции ТТЕСТ(), считая, что в двух группах одни и те же пациенты (данные до принятия лекарства и после).

4. С помощью соответствующей процедуры пакета Анализ данных вычислить $Р_{Н_0}$, учитывая, что в группах для одни и те же пациенты. Сравнить результаты пунктов 3 и 4. На основании полученных результатов записать в тетрадь вывод об эффективности нового лекарственного средства.

5. Выявить эффективность нового лекарственного средства, применив повторно функцию ТТЕСТ(), считая, что в двух группах разные пациенты (тип 3).

6. С помощью соответствующей процедуры пакета Анализ данных вычислить $Р_{Н_0}$, для разных пациентов в группах. Сравнить результаты пунктов 5 и 6. На основании полученных результатов записать в тетрадь вывод об эффективности нового лекарственного средства в данном случае.

7. Используя значения доверительных интервалов для среднего,

предположив, что в обеих группах нормальное частотное распределение, проверить наличие достоверности различий в двух группах (пациенты одни и те же).

8. Проверить полученные результаты с помощью функции ХИ2ТЕСТ().

9. На основании результатов описательной статистики обоснуйте вывод: применение какой из ранее указанных функций является правомочным в данном случае.

ХОД ВЫПОЛНЕНИЯ РАБОТЫ

Изучение функций ТТЕСТ() и ХИ2ТЕСТ()

1. Используя материалы встроенной программной справочной системы Microsoft Excel, найдите и запишите в тетрадь в таблицу 7 основную информацию о встроенных функциях, выявляющих достоверность различий: ТТЕСТ() и ХИ2ТЕСТ().

Таблица 7.

Информация о встроенных функциях

Название функции	Назначение функции	Синтаксис функции	Примеры применения
ТТЕСТ()			
ХИ2ТЕСТ()			

Параметрические методы выявления достоверности различий

Выявление эффективности нового лекарственного средства с помощью парного теста критерия Стьюдента (одни и те же пациенты)

1. Скопируйте из папки «Z:\ Материалы для работы\Статистика» в свою папку файл *Пр.зан.№4-Стьюдент.xls*, сохраните его с именем *Пр.зан.№4- Иванов А. — 24 леч,* подставив свою фамилию, номер группы.

2. Получите на Рабочем листе 1:
   - ***описательные статистики*** для приведенных групп. Результаты (значения среднего, медианы, моды, уровня надежности) запишите в таблицу 8;
   - вычислите ***границы доверительных интервалов*** для каждой группы, определите, есть ли перекрытие доверительных интервалов, т.е. наличие общего среднего в двух группах, результат внесите в таблицу 8;
   - в таблице 8 приведите ***вывод о нормальности частотных распределений в каждой группе*** и применяемом методе статистического анализа (параметрический, непараметрический).

Таблица 8.

Результаты описательной статистики

Статистика	**Выборка**		**Вывод о нормальности в группе (да/нет)**	
	контрольная	**исследуемая**	**контрольная**	**исследуемая**
Среднее				
Медиана				
Мода				
Уровень надежности				
Метод статистического анализа (параметрический, непараметрический).				
Доверительный интервал			**Есть пересечение ДИ, т.е. наличие общего среднего? (да / нет)**	
Нижняя граница ДИ				
Верхняя граница ДИ				

3. Вычислите на Рабочем листе 1 вероятность достоверности различия двух выборок с помощью функции ТТЕСТ(), Тип=1, Хвосты=2. Результат запишите в тетрадь, в таблицу 9.

4. Вычислите вероятность нулевой гипотезы с помощью процедуры «Парный двухвыборочный t-тест для средних», установив параметры окна в соответствии с рисунком 42. Запишите значение ***P(T<=t) двухстороннее*** в таблицу 9.

5. Сравните его со значением, полученным ранее с помощью функции ТТЕСТ().

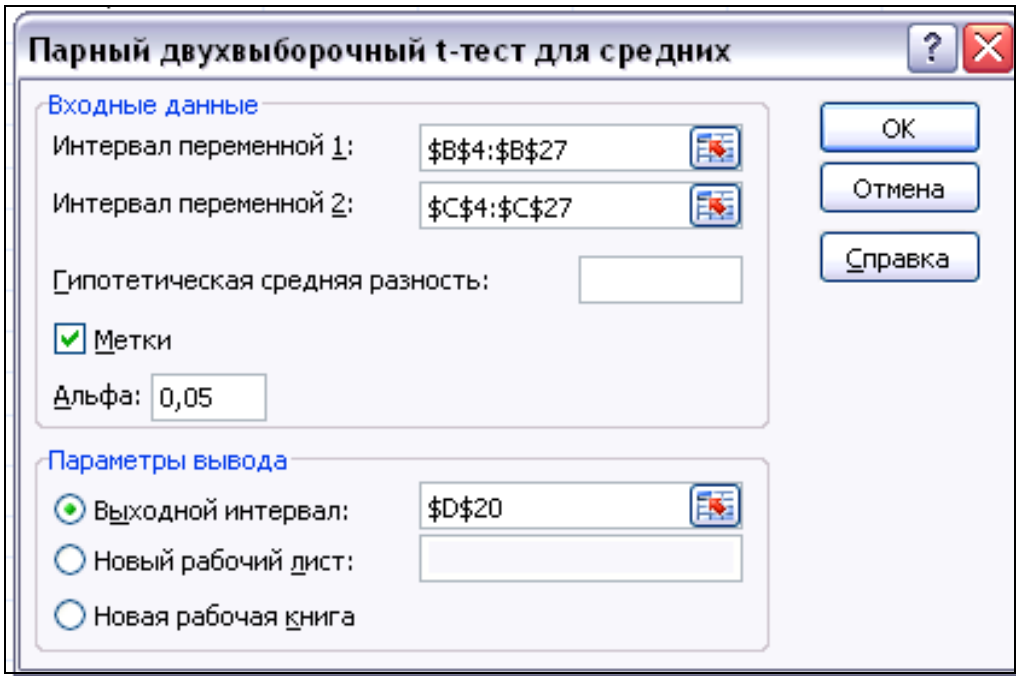


Рис. 42. Вид окна процедуры ***«Парный двухвыборочный t-тест для средних»***

6. В тетради запишите аргументированный вывод об эффективности нового лекарственного средства, используя полученный результат и значения средних в группах. Заполните в таблице 9, соответствующие значения.

Таблица 9.

Результаты выполнения критериев Стьюдента и Пирсона

Критерий	**PH_0**	**Принимаемая гипотеза (H_0 / H_1)**	**Наличие достоверных различий (*да/нет*)**	**Эффективность нового фарм. препарата? (*да/нет*)**	**Правомочно применение метода (*да/нет*)**
Стьюдента (ТТЕСТ()) тип=1					
Стьюдента (ТТЕСТ()) тип=3					
Парный t-тест для средних (Пакет анализа)					
Двухвыборочный t-тест с разными дисперсиями (Пакет анализа)					
Согласия ПИРСОНА (ХИ2ТЕСТ)					

Выявление достоверности различий признака в выборках, используя критерий Стьюдента — пациенты разные

1. Примените функцию ТТЕСТ() для вычисления эффективности лекарственного средства в случае, когда в группах разные пациенты (Тип=3, Хвосты 2). Результат запишите в тетрадь в таблицу 9.

2. Вычислите вероятность нулевой гипотезы с помощью процедуры ***«Двухвыборочный t-тест с различными дисперсиями»,*** установив в окне процедуры адрес ***Выходного интервала — D35.***

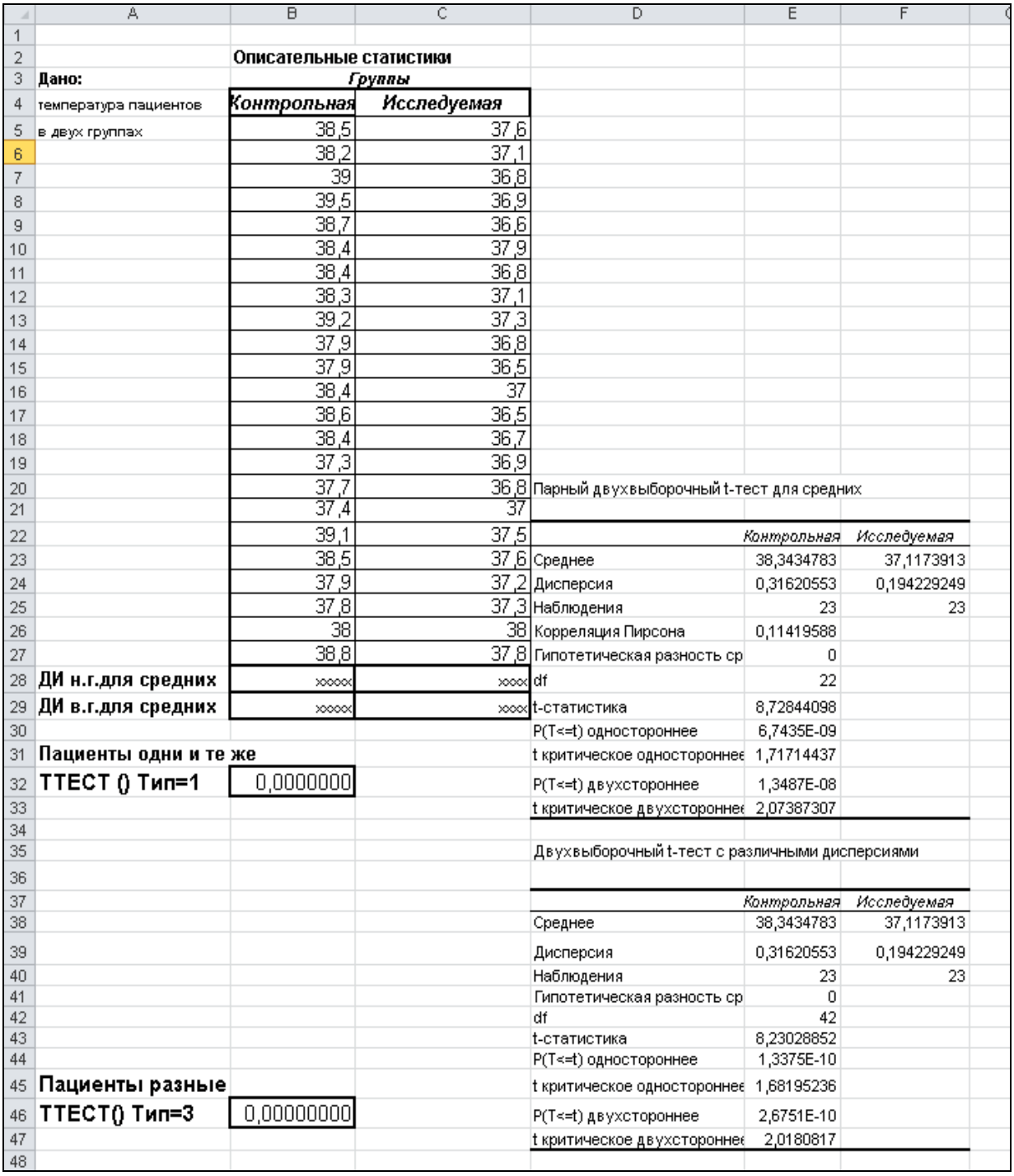
Запишите значение ***P(T<=t) двухстороннее*** в тетрадь в таблицу 9. Сравните его со значением, полученным с помощью функции ТТЕСТ().

3. Запишите в тетради подробный аргументированный вывод об эффективности нового лекарственного средства, используя полученный результат и значения средних в группах.

Общий вид Рабочего листа 1 представлен на рисунке 43.

4. Запишите в тетрадь вывод о том, как отличаются результаты, полученные разными методами. Какой метод более рационален?

5. Сохраните результаты работы в Вашем файле.



	A	B	C	D	E	F
2		**Описательные статистики**				
3	**Дано:**		***Группы***			
4	температура пациентов	***Контрольная***	***Исследуемая***			
5	в двух группах	38,5	37,6			
6		38,2	37,1			
7		39	36,8			
8		39,5	36,9			
9		38,7	36,6			
10		38,4	37,9			
11		38,4	36,8			
12		38,3	37,1			
13		39,2	37,3			
14		37,9	36,8			
15		37,9	36,5			
16		38,4	37			
17		38,6	36,5			
18		38,4	36,7			
19		37,3	36,9			
20		37,7	36,8	Парный двухвыборочный t-тест для средних		
21		37,4	37			
22		39,1	37,5		*Контрольная*	*Исследуемая*
23		38,5	37,6	Среднее	38,3434783	37,1173913
24		37,9	37,2	Дисперсия	0,31620553	0,194229249
25		37,8	37,3	Наблюдения	23	23
26		38	38	Корреляция Пирсона	0,11419588	
27		38,8	37,8	Гипотетическая разность ср	0	
28	**ДИ н.г.для средних**	×××××	×××××	df	22	
29	**ДИ в.г.для средних**	×××××	×××××	t-статистика	8,72844098	
30				P(T<=t) одностороннее	6,7435E-09	
31	**Пациенты одни и те же**			t критическое одностороннее	1,71714437	
32	**ТТЕСТ () Тип=1**	0,0000000		P(T<=t) двухстороннее	1,3487E-08	
33				t критическое двухстороннее	2,07387307	
34						
35				Двухвыборочный t-тест с различными дисперсиями		
36						
37					*Контрольная*	*Исследуемая*
38				Среднее	38,3434783	37,1173913
39				Дисперсия	0,31620553	0,194229249
40				Наблюдения	23	23
41				Гипотетическая разность ср	0	
42				df	42	
43				t-статистика	8,23028852	
44				P(T<=t) одностороннее	1,3375E-10	
45	**Пациенты разные**			t критическое одностороннее	1,68195236	
46	**ТТЕСТ() Тип=3**	0,00000000		P(T<=t) двухстороннее	2,6751E-10	
47				t критическое двухстороннее	2,0180817	
48						

Рис. 43. Вид Рабочего листа 1 с результатами выполнения задания

Выявление достоверности различий признака в выборках с нормальным распределением по доверительным интервалам среднего

1. По данным таблицы 8 (наличие пересечения доверительных интервалов) сделайте вывод, могут ли эти группы иметь общее среднее, и как следствие есть ли между ними статистически значимое различие?

2. Сделайте выводы, о том какой способ выявления достоверности различий при нормальном частотном распределении более рационален, запишите его в тетрадь.

Непараметрические методы выявления достоверности различий

Выявление достоверности различий признака в выборках с любым частотным распределением с помощью критерия Пирсона χ^2

Вычислить достоверность различия выборок с любым частотным распределением можно с помощью критерия Пирсона χ^2, *который реализует функция ХИ2ТЕСТ()*. Для того, чтобы применить эту функцию необходимо:

1) от признака перейти к его частотам (см. практическую работу №2);

2) вычислить ожидаемые значения частот.

Вычисление частот

1. Перейдите в Вашем файле на Рабочий лист 2. На этом листе размещены частоты, полученные Вами для построения гистограмм по двум группам.

Учитывая, что элементарная частота при применении функции ХИ2ТЕСТ() должна быть не меньше 3, эти диапазоны следует укрупнить (необязательно, чтобы они были одинаковыми);

2. Преобразуйте исходную таблицу так, чтобы элементарная частота в каждом диапазоне была не меньше 3:

- введите в строках 17-19 электронной таблицы соответствующий текст и количество, посчитав недостающие данные самостоятельно.

3. Расчет ***ожидаемых*** значений проводите по следующему алгоритму:

- введите в ячейки C23:D23 формулы для вычисления общего количества объектов в каждой из групп;
- введите в ячейки E20:E23 формулы для вычисления суммы частот обеих групп на каждом диапазоне значений признака;
- введите формулы в ячейки F20:F23 для вычисления доли признака в обеих группах на заданном диапазоне;
- введите в ячейки G20:H23 формулы для вычисления ожидаемых значений для каждой из групп, как произведение доли признака на общее количество объектов в конкретной группе.

	A	B	C	D	E	F	G	H
1								
2		Критерий ХИ-квадрат						
3								
4			Диапазоны	Частота				
5			значений	*Контрольная*	*Исследуемая*			
6			36,5 -37	2				
7			37,1 - 37,5	2				
8			37,6 - 38	6				
9			38,1 - 38,5	7				
10			38,6 - 39	4				
11			39,1 -39,5	2				
12			39,6- 40	0				
13				=СУММ(D6:D12)	=СУММ(E6:E12)			
14								
15								
16								
17		Укрупнение диапазонов			Расчет ожидаемых значений			
18		Диапазоны	Частота				Ожидаемые значения	
19		значений	*Контрольная*	*Исследуемая*	Сумма	% признака	*Контрольная*	*Исследуемая*
20		36,5 -37,5			=СУММ(C20:D20)	=E20/$E$23	=F20*$C$23	=F20*$D$23
21		37,6 - 38			=СУММ(C21:D21)	=E21/$E$23	=F21*$C$23	=F21*$D$23
22		38,1 - 39,5			=СУММ(C22:D22)	=E22/$E$23	=F22*$C$23	=F22*$D$23
23		Итого	=СУММ(C20:C22)	=СУММ(D20:D22)	=СУММ(E20:E22)	=E23/$E$23	=F23*$C$23	=F23*$D$23
24								
25	ХИ2ТЕСТ()=							

Рис. 44. Вид Рабочего листа с формулами для вычисления ожидаемых значений

4. Вычислите значение ХИ2ТЕСТ() в ячейке В25, заполнив параметры функции в соответствии с рисунком 45.

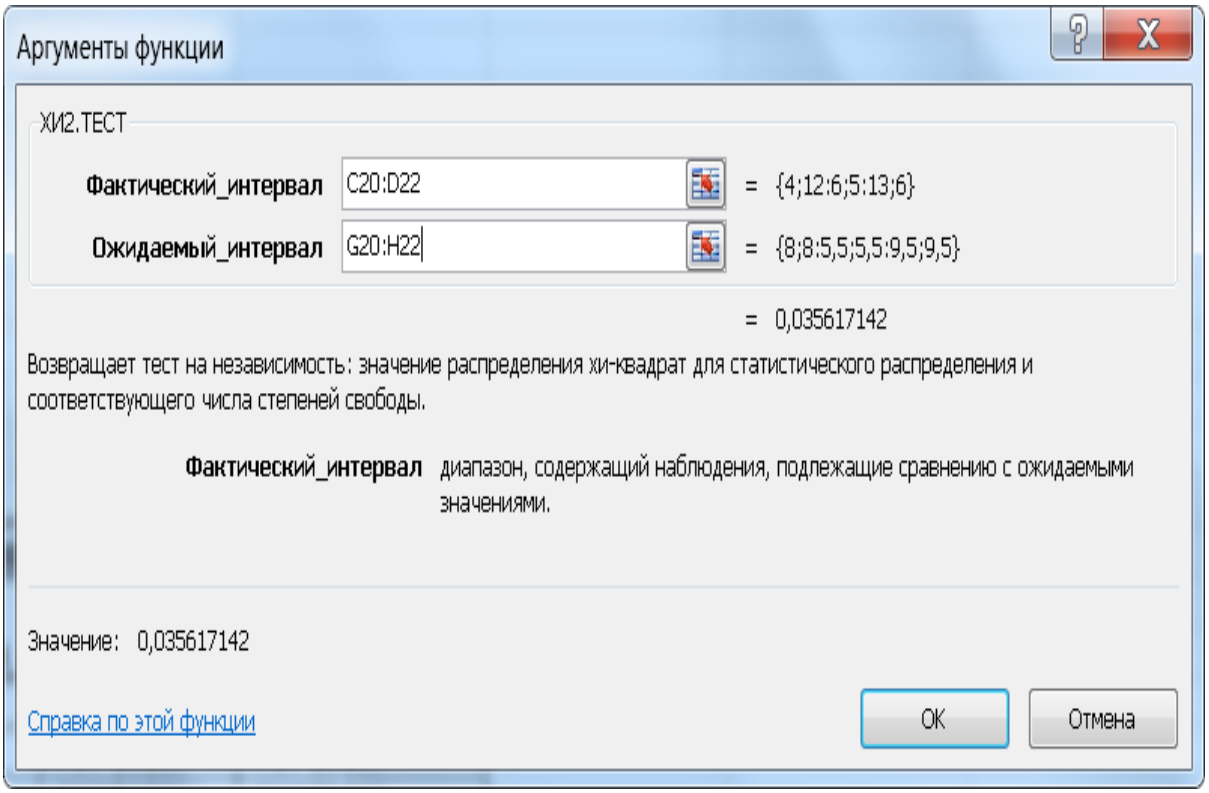


Рис. 45. Параметры функции ХИ2ТЕСТ()

5. Запишите полученный результат в тетрадь в таблицу 9, сохраните свой файл с результатами работы.

6. На основании полученного значения функции ХИ2ТЕСТ() сделайте вывод, есть ли статистически значимые различия в виде частотных распределений в двух группах.

7. Сравните значения температур в двух группах, в какой группе частоты ближе к норме? Сделайте аргументированный вывод об эффективности нового лекарственного средства, запишите его в тетрадь.

8. Сравните полученный вывод с результатами анализа по критерию Стьюдента.

9. Какой метод правомочно использовать на заданных выборках? Ответ аргументируйте и запишите в тетрадь.

Вопросы для самоконтроля

1. В каких случаях для статистического анализа медико-биологических данных используются параметрические методы и непараметрические?

2. Какие виды гипотез выдвигаются при выявлении достоверности различий?

3. Как формулируется обычно нулевая гипотеза при выявлении достоверности различий параметрическими методами?

4. Как формулируется нулевая гипотеза при проведении статистического анализа данных непараметрическими методами?

5. Что такое уровень значимости? Его математическая интерпретация.

6. При каком значении $P_{но}$ считается справедливой нулевая гипотеза (H_o)?

7. Всегда ли при справедливости H_o можно считать, что инновация неэффективна?

8. Всегда ли при справедливости H_1 можно считать, что инновация эффективна?

9. Приведите алгоритм определения справедливости утверждения об эффективности нового фармацевтического препарата.

10. Приведите синтаксис функции, реализующей критерий Стьюдента и значения всех его параметров.

11. Сущность критерия χ^2 Пирсона. Как этот критерий реализуется в Excel? Назовите достоинства и ограничения при применении этого критерия. Как интерпретируются полученные результаты?

12. Назовите ограничения на применение критерия Стьюдента.

13. Назовите ограничения на применение критерия χ^2 Пирсона.

14. Назовите процедуры пакета Анализ данных, реализующие критерий Стьюдента.

15. В чем сущность процедуры «Парный двухвыборочный t-тест для средних», какому значению параметра «Тип» функции ТТЕСТ() она соответствует?

16. В чем сущность процедуры «Двухвыборочный t-тест с различными дисперсиями», какому значению параметра «Тип» функции ТТЕСТ() она соответствует?

17. Запишите синтаксис функции ХИ2ТЕСТ.

18. Приведите алгоритм вычисления параметра «ожидаемый интервал» при использовании функции ХИ2ТЕСТ.

5. МЕТОДЫ ВЫЯВЛЕНИЯ ВЗАИМОСВЯЗЕЙ

5.1. ТЕОРЕТИЧЕСКИЕ ОСНОВЫ

Понятие корреляционной зависимости

Все явления в мире взаимосвязаны. Связь, при которой изменение одного признака влечет изменение распределения другого, называется статистической связью. Взаимосвязь между явлениями может быть качественная и количественная. Качественную взаимосвязь позволяет выявить *корреляция*, а количественную — *регрессия* (на лечебном факультете эту тему изучают на 6 курсе). Если для определения взаимосвязей используются средние величины, то такие критерии называют критериями корреляционной связи.

Корреляционная зависимость может быть *линейной, нелинейной*, *прямой* и *обратной*. Примерами прямой и обратной связи могут быть следующие зависимости. При повышении температуры тела повышается частота дыхания — прямая зависимость. Увеличение количества прививок приводит к уменьшению количества заболеваний пациентов — обратная зависимость.

Параметр, характеризующий степень линейной взаимосвязи между выборками, называется *коэффициентом корреляции.*

Значение коэффициента корреляции изменяется от –1 (строгая обратная линейная зависимость) до 1 (строгая прямая пропорциональная зависимость). При значении 0 — линейной связи между выборками нет. Для оценки степени взаимосвязи руководствуются следующими правилами. Если абсолютная величина значения коэффициента корреляции (***r***) больше, чем **0,95**, то между параметрами существует практически линейная зависимость. Если абсолютная величина коэффициент корреляции находится в диапазоне от **0,8 до 0, 95**, то говорят о *сильной степени связи* между параметрами; при **0,6 < r < 0,8** — говорят о наличии *средней степени связи* между параметрами; **0,4 < r < 0,6** — *умеренной;* при **r < 0,4** — считают, что взаимосвязь между параметрами *слабая.*

Параметрическая корреляция Пирсона

При нормальных частотных распределениях в выборках используется параметрическая корреляция Пирсона. Значение коэффициента линейной параметрической корреляции позволяет получить встроенная функция Microsoft Excel **КОРРЕЛ (M1,M2),**

где **M1** — массив первой выборки, **M2** — массив второй выборки. Пример применения функции КОРРЕЛ() представлен на рисунке 46**.**

В приведенном примере результат выполнения функции КОРРЕЛ() позволяет оценить силу связи между частотой сердечных сокращений и частотой дыхания при исследуемой патологии. В качестве параметров функции используются диапазоны ячеек *B4:B10*, в которых расположены данные о частоте сердечных сокращений, и *C4:C10*, где находятся данные о частоте дыхания. Полученное значение коэффициента корреляции **r**=0,8537, свиде-

тельствует о прямой сильной зависимости между частотой сердечных сокращений и частотой дыхания.

B11 =КОРРЕЛ(B4:B10;C4:C10)

	A	B	C	D	E	F	G	H
1		Выявление взаимосвязи между частотой сердечных сокращений						
2		и частотой дыхания при исследуемой патологии						
3		ЧСС	ЧД					
4		120	20					
5		84	15					
6		134	18					
7		92	16					
8		113	19					
9		90	16					
10		80	15					
11	КОРРЕЛ	0,853767						

Рис. 46. Вид фрагмента Рабочего листа Excel с функцией КОРРЕЛ( )

Корреляцию Пирсона можно также рассчитать с помощью процедуры «***Корреляция»*** пакета Анализ данных (рис. 47).

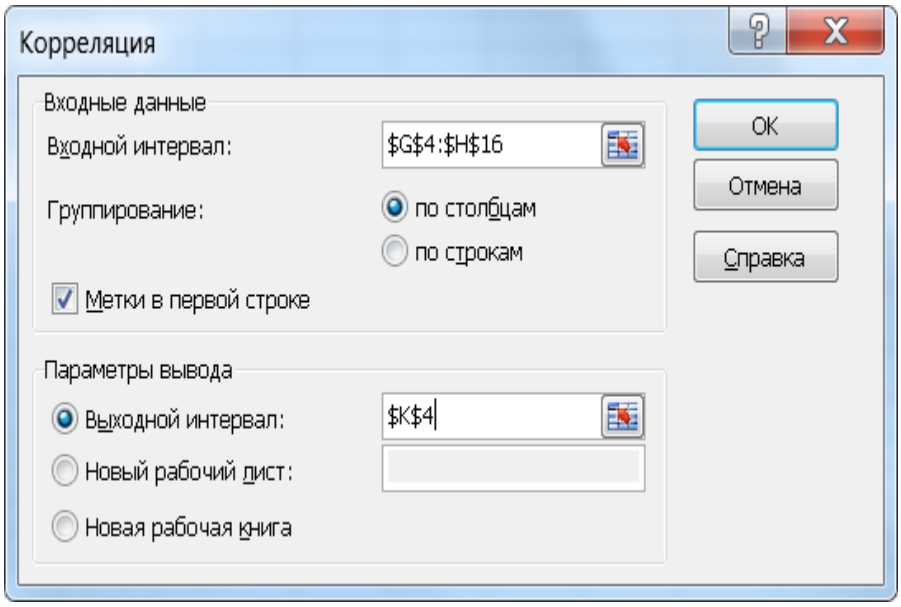


Рис. 47. Вид окна процедуры «Корреляция» пакета Анализ данных

Результатом работы этой процедуры является корреляционная матрица, в которой по строкам и столбцам располагаются исследуемые признаки (переменные). В корреляционной матрице на пересечении строк и столбцов расположены коэффициенты корреляции этих признаков.

Например, требуется определить наличие взаимосвязи между числом ясных дней, количеством посещения массажа и водного лечения. Вид корреляционной матрицы, полученной с помощью процедуры «Корреляция» пакета Анализ данных, представлен на рисунке 48. Коэффициенты корреляции обведены овалами.

	Число ясных дней	*Количество посещений массажа*	*Количество посещений водного лечения*
Число ясных дней	1		
Количество посещений массажа	-0,111045377	1	
Количество посещений водного лечения	0,974575588	-0,018443098	1

Рис.48. Вид корреляционной матрицы

Непараметрическая ранговая корреляция Спирмена

В тех случаях, когда законы частотного распределения в выборках с анализируемыми признаками не являются нормальными, для оценки силы связи применяют непараметрические методы. Одним из методов выявления силы связи является вычисление рангового коэффициента корреляции Спирмена, который позволяет найти коэффициент ранговой корреляции (ρ_{xy}). В Microsoft Excel нет функции, реализующей получение этого коэффициента, но его можно вычислить, зная соответствующий алгоритм.

Коэффициент ранговой корреляции вычисляется по формуле:

$$\rho_{xy} = 1 - 6*\sum d_i^2/(n*(n^2-1)) \quad (1)$$

где d_i — разность рангов пары значений, соответствующих одному объекту (номеру одной точки в таблице рангов);

n — количество объектов в выборке.

Краткий алгоритм вычисления коэффициента ранговой корреляции Спирмена:

1. Определить ранги пары исследуемых признаков (каждого объекта).
2. Рассчитать разность рангов для каждого объекта (d_i).
3. Вычислить квадраты рангов (d_i^2).
4. Рассчитать коэффициент ранговой корреляции по формуле (1).

Реализация алгоритма вычисления коэффициента ранговой корреляции Спирмена в среде электронных таблиц Microsoft Excel

Шаг 1. Вычислить ранги для значений каждой пары (точки) исследуемых признаков.

Ранги следует вычислять с помощью процедуры «Ранг и персентиль» пакета Анализ данных. Процедура «Ранг и персентиль» самому большому значению признака присваивает самый низкий ранг, а самому маленькому значению — самый высокий ранг.

Например, есть следующие данные: возраст студента группы и его средний балл (рис. 49). Следует определить, есть ли зависимость между воз-

растом и успеваемостью студента.

Возраст студента группы	Средний балл студента
18	7,7
19	8,5
20	8,6
18	6,6
22	6,9
23	6,6

Рис. 49. Пример данных для выявления корреляционной связи по Спирмену

Применив для массива данных в ячейках C2:D8 процедуру ***«Ранг и персентиль»***, получим таблицу рангов (рис. 50).

Возраст студента группы	Средний балл студента	Точка	Возраст студента группы	Ранг	Процент	Точка	Средний балл студента	Ранг	Процент
18	7,7	6	23	1	100,00%	3	8,6	1	100,00%
19	8,5	5	22	2	80,00%	2	8,5	2	80,00%
20	8,6	3	20	3	60,00%	1	7,7	3	60,00%
18	6,6	2	19	4	40,00%	5	6,9	4	40,00%
22	6,9	1	18	5	0,00%	4	6,6	5	0,00%
23	6,6	4	18	5	0,00%	6	6,6	5	0,00%

Рис. 50. Ранги исследуемых признаков, полученные с помощью процедуры ***«Ранг и персентиль»***

Переформируем эти данные, поставив каждой точке (объекту) в соответствие ее пару рангов. Для этого:

1) отсортируем массив E2:H8 по возрастанию поля «Точка»;

2) отсортируем массив I2:L8 по возрастанию поля «Точка»;

3) в ячейки строки 11 введем соответствующие текстовые значения (рис. 51.);

4) скопируем соответствующие признакам ранги из ячеек G3:G8 в ячейки C12:C17 и из ячеек K3:K8 в ячейки D12:D17.

	Возраст студента группы	Средний балл студента	Точка	Возраст студента группы	Ранг	Процент	Точка	Средний балл студента	Ранг	Процент
3	18	7,7	1	18	5	0,00%	1	7,7	3	60,00%
4	19	8,5	2	19	4	40,00%	2	8,5	2	80,00%
5	20	8,6	3	20	3	60,00%	3	8,6	1	100,00%
6	18	6,6	4	18	5	0,00%	4	6,6	5	0,00%
7	22	6,9	5	22	2	80,00%	5	6,9	4	40,00%
8	23	6,6	6	23	1	100,00%	6	6,6	5	0,00%

Точка	Ранг возраста студентов	Ранг среднего балла студентов	d	d^2
1	5	3	2	4
2	4	2	2	4
3	3	1	2	4
4	5	5	0	0
5	2	4	-2	4
6	1	5	-4	16
			Сумма=	32

r = 0,086

Рис. 51. Вид Рабочего листа с результатами вычислений

Шаг 2. Вычислить разности соответствующих рангов (d).

Шаг 3. Возвести разности рангов (d) в квадрат.

Шаг 4. Вычислить сумму квадратов разности.

Шаг 5. Вычислить коэффициент ранговой корреляции Спирмена, применив формулу 1.

Вид таблицы Microsoft Excel с формулами приведен на рисунке 52. Результаты вычислений приведены на рисунке 51.

	Точка	Ранг возраста студентов группы	Ранг среднего балла студентов	Разность рангов (d)	d^2
3	6	1	5	=C3-D3	=E3*E3
4	5	2	4	=C4-D4	=E4*E4
5	3	3	1	=C5-D5	=E5*E5
6	2	4	2	=C6-D6	=E6*E6
7	1	5	3	=C7-D7	=E7*E7
8	4	5	5	=C8-D8	=E8*E8
9				Сумма	=СУММ(F3:F8)
11				Коэффициент ранговой корреляции Спирмена	=1- 6*F9/(6*35)

Рис. 52. Формулы для вычисления коэффициента ранговой корреляции Спирмена

Полученный результат, значение коэффициента ранговой корреляции ρ_{xy}**=0,086**, свидетельствуют о том, что связь между возрастом и средним баллом студентов группы незначительна.

5.2. ПРИМЕНЕНИЕ ПАРАМЕТРИЧЕСКОГО И НЕПАРАМЕТРИЧЕСКОГО КОРРЕЛЯЦИОННОГО АНАЛИЗА ДАННЫХ

ЦЕЛИ ЗАНЯТИЯ

1. Ознакомиться с основными понятиями выявления взаимосвязей (параметрическая корреляция Пирсона, непараметрическая корреляция Спирмена).
2. Получить навыки применения встроенной функции Microsoft Excel, вычисляющей коэффициент корреляции Пирсона.
3. Научиться формулировать выводы о наличии качественной взаимосвязи в двух выборках.
4. Овладеть навыками применения встроенной функции КОРРЕЛ() и процедуры «Корреляция» пакета Анализ данных, реализующей вычисление корреляции Пирсона.
5. Ознакомиться с алгоритмом вычисления непараметрического рангового коэффициента корреляции Спирмена.

МЕТОДИКА ВЫПОЛНЕНИЯ РАБОТЫ

Постановка задачи 1

Дано: ***С***ведения о распределении водителей автобусов по возрастным группам по двум автобазам ***(***таблица 10)

Таблица 10.

Сведения о распределении работников двух автобаз по возрастным группам

Возрастная группа	Численность работающих	
	Автобаза №1	Автобаза №2
до 20 лет	150	250
21 - 39	420	450
40 - 59	350	100
60 и старше	100	45
Всего	**1020**	**845**

Требуется:

1. Построить с помощью Мастера диаграмм Microsoft Excel гистограмму частотных распределений водителей по возрастным группам для двух автобаз.
2. Обосновать и записать в тетрадь выбор методов дальнейшей статистической обработки данных: параметрический, непараметрический (по описательным статистикам);
3. С помощью критерия Стьюдента выявить наличие статистически значимых различий в этих выборках.
4. С помощью функции ХИ2ТЕСТ() выявить наличие статистически значимых различий распределения водителей по возрастным группам на двух автобазах. Вывод записать в тетрадь.
5. В тетради записать обоснование, какой метод следовало применять

в данном случае и почему.

Постановка задачи 2

Дано: Рост и масса учащихся одного из младших классов средней школы в таблице11.

Таблица 11.

Сведения о росте и массе учащихся младшего школьного возраста одного из классов средней школы

Порядковый номер	Рост тела (см) (X)	Масса тела (кг) (У)
1	110	20
2	112	22
3	120	24
4	127	25
5	130	25
6	135	27
7	135	25
8	140	30
9	145	35
10	145	37

Требуется:

1. Вычислить описательные статистики по каждой выборке. Сформулировать выводы о нормальности частотных распределений в этих группах, записать их в тетрадь. Обосновать выбор метода дальнейшего анализа (параметрический, непараметрический).
2. С помощью функции КОРРЕЛ() вычислить наличие качественной взаимосвязи между ростом и массой ученика младшего школьного возраста одного из классов средней школы.
3. С помощью процедуры «Корреляция» пакета Анализ данных проверить полученное значение.
4. Сделать вывод о направлении и силе взаимосвязи между исследуемыми признаками.
5. Используя представленный выше алгоритм, вычислить коэффициент ранговой корреляции Спирмена.
6. В тетради (в таблицу 13) записать:
   - значение среднего, медианы, моды для каждого признака (рост тела, масса тела);
   - тип частотного распределения для каждого признака (нормальное, ненормальное);
   - выбранный метод статистической обработки данных (параметрический, непараметрический), обосновав его применение;
   - синтаксис примененной функции КОРРЕЛ() из строки формул и полученное значение коэффициента корреляции Пирсона с помощью этой функции;
   - команду вызова *процедуры **«Корреляция»*** из пакета Анализ дан-

ных;

– значение коэффициента корреляции, полученное с помощью этой процедуры;
– алгоритм вычисления коэффициента ранговой корреляции Спирмена, формулу для вычисления этого коэффициента;
– значение коэффициента ранговой корреляции Спирмена;
– вывод о наличии и силе связи между ростом и массой тела.

7. Сравнить значения двух коэффициентов корреляции, обосновать какой метод следовало использовать в данном случае.

ХОД РАБОТЫ

Рекомендации по решению задачи 1

1. Скопируйте из папки «Z:\ Материалы для работы\Статистика» в свою папку файл *Пр.зан.№5-Корреляция.xls*, сохраните его с именем *Пр.зан.№5- Иванов А. — 24 леч,* подставив свою фамилию.
2. Выполняйте задание на Рабочем листе с именем *«Задание 1»*.
3. Разместите результаты задания 1 на соответствующем Рабочем листе, общий вид которого представлен на рисунке 53.

✓ **Обратите внимание!** При применении функции ХИ2ТЕСТ() в связи с тем, что исходные данные представлены частотами, переходить от признака к частотам не следует. По заданным исходным частотам необходимо рассчитать ожидаемые значения, используя формулы, приведенные в практической работе №4.

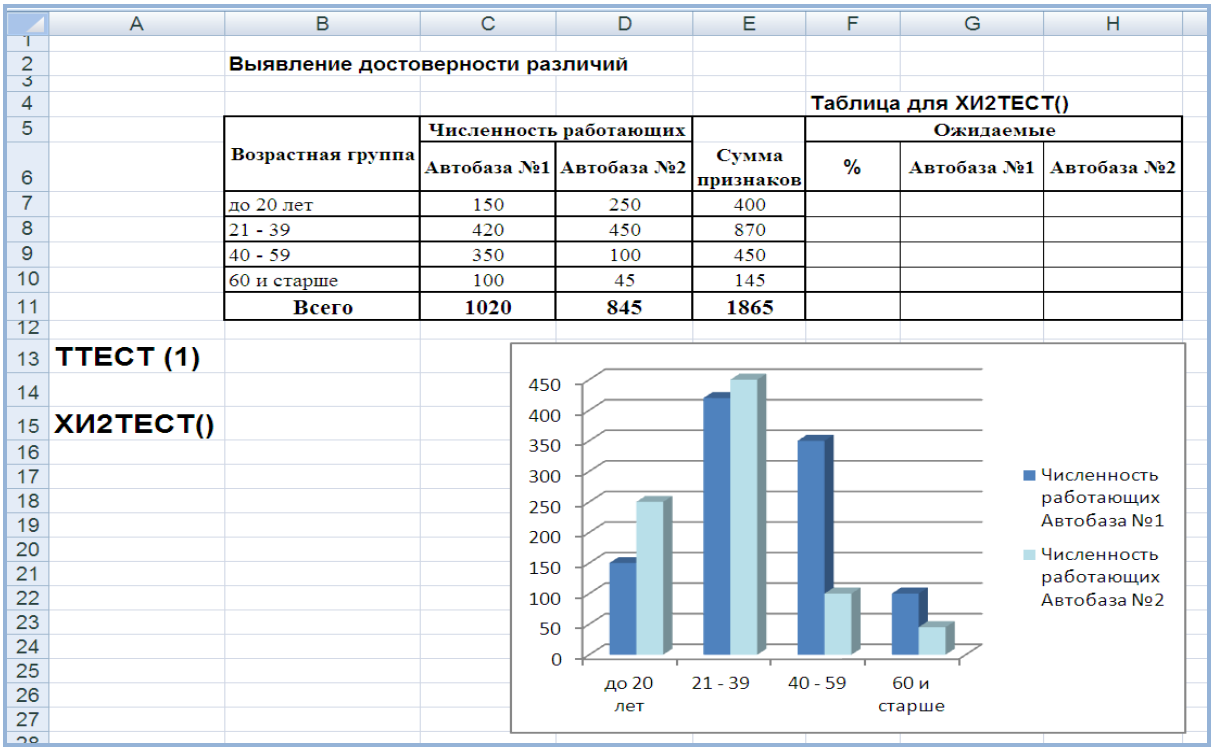


Рис. 53. Вид Рабочего листа с результатами выполнения задания 1

Решение задачи 2

Выявление нормальности частотного распределения

1. На Рабочем листе *«Пирсон»* вычислите описательные статистики по двум выборкам.

2. Используя, один из критериев нормальности частотного распределения сделайте вывод о нормальности распределений признаков: масса тела и рост, запишите его в тетрадь в таблицу 13.

3. Сделайте вывод о том, какие методы вычисления корреляции следует применять в данном случае и почему, запишите его в тетрадь в таблицу 13.

Выявления корреляции параметрическими методами

Определение взаимосвязей между признаками с помощью корреляции Пирсона

1. Изучите синтаксис и назначение функции КОРРЕЛ().

Используя материалы встроенной программной справочной системы Microsoft Excel, найдите и запишите в тетрадь в таблицу 12 основную информацию о встроенной функции КОРРЕЛ(), вычисляющей коэффициент корреляции Пирсона.

Таблица 12.

Информация о встроенной функции

Название функции	Назначение функции	Синтаксис функции	Примеры применения
КОРРЕЛ()			

2. На Рабочем листе с именем *«Пирсон»* вычислите значение коэффициента корреляции Пирсона, выявляющим направление и силу связи между ростом и массой учащихся младших классов. Результат запишите в тетрадь (таблица 13).

Таблица 13.

Результаты определения силы взаимосвязи между ростом тела и его массой

Наименование признака	Среднее	Медиана	Мода	Нормальность	Метод обработки	Синтаксис, значение r Пирсона	r пакета Анализ данных	r Спирмена
Рост тела								
Масса тела								

Команда вызова процедуры «Корреляция» _____

Алгоритм вычисления коэффициента ранговой корреляции Спирмена:

_____

_____

_____

Формула _____

Вывод_____

3. Проверьте полученное значение, применив процедуру **«Корреляция»** пакета Анализ данных. Разместите результаты работы в соответствии с рисунком 54.

4. Запишите в тетрадь аргументированный вывод о наличии связи между ростом и массой учащихся и ее напрвлении и силе.

	A	B	C	D	E	F	G	H	I
1		Макет для вычисления коэффициента ранговой корреляции Пирсона							
2		Исходные данные							
3		Порядковый номер	Рост тела (см) (X)	Масса тела (кг) (У)		*Рост тела (см) (X)*		*Масса тела (кг) (У)*	
4		1	110	20					
5		2	112	22		Среднее	130	Среднее	27
6		3	120	24		Стандартная ошибка	3,99	Стандартная ошибка	1,7256239
7		4	127	25		Медиана	133	Медиана	25
8		5	130	25		Мода	135	Мода	25
9		6	135	27		Стандартное отклонение	12,62	Стандартное отклонение	5,46
10		7	135	25		Дисперсия выборки	159,21	Дисперсия выборки	29,78
11		8	140	30		Эксцесс	-1,03	Эксцесс	-0,07
12		9	145	35		Асимметричность	-0,44	Асимметричность	0,87
13		10	145	37		Интервал	35	Интервал	17
14						Минимум	110	Минимум	20
15						Максимум	145	Максимум	37
16						Сумма	1299	Сумма	270
17						Счет	10	Счет	10
18						Наибольший(1)	145	Наибольший(1)	37
19						Наименьший(1)	110	Наименьший(1)	20
20						Уровень надежности(95,0%)	9,03	Уровень надежности(95,0%)	3,90
21									
22					*Рост тела (см) (X)*	*Масса тела (кг) (У)*			
23				Рост тела (см) (X)	1				
24	КОРРЕЛ()=			Масса тела (кг) (У)		1			

Рис. 54. Вид Рабочего листа *«Пирсон»* с результатами работы

Выявления взаимосвязей непараметрическим методом *(Коэффициент ранговой корреляции Спирмена)*

1. На Рабочем листе *«Спирмен»*, следуя алгоритму, приведенному в теоретическом блоке, вычислите *ранги признаков* для каждого субъекта с помощью процедуры «Ранг и персентиль», *их разность*, *квадраты разности*, *сумму квадратов* и по формуле (1) вычислите *коэффициент ранговой корреляции Спирмена*. Результаты разместите на Рабочем листе в соответствии с рисунком 55.

2. Запишите в тетрадь вывод о направлении и силе связи между ростом и весом ученика.

3. Сравните результаты с ранее полученными параметрическим методом.

4. Применение какого метода является более корректным в данном случае?

5. Подтверждается ли вывод о наличии связи между ростом и весом учащихся младшего школьного возраста?

	A	B	C	D	E	F	G	I	J	K	L	M	N	O	P
1	Макет для вычисления коэффициента ранговой корреляции Спирмена														
2	Исходные данные			Результаты											
3				Ранги			Квадрат								
4	Порядковый номер	Рост тела (см) (X)	Масса тела (кг) (У)	Рост тела (см) (X)	Масса тела (кг) (У)	Разности рангов (d)	разности рангов (d^2)	Точка	Рост тела (см) (X)	Ранг	Процент	Точка	Масса тела (кг) (У)	Ранг	Процент
5	1	110	20					9	145	1	88,80%	10	37	1	100,00%
6	2	112	22					10	145	1	88,80%	9	35	2	88,80%
7	3	120	24					8	140	3	77,70%	8	30	3	77,70%
8	4	127	25					6	135	4	55,50%	6	27	4	66,60%
9	5	130	25					7	135	4	55,50%	4	25	5	33,30%
10	6	135	27					5	130	6	44,40%	5	25	5	33,30%
11	7	135	25					4	127	7	33,30%	7	25	5	33,30%
12	8	140	30					3	120	8	22,20%	3	24	8	22,20%
13	9	145	35					2	112	9	11,10%	2	22	9	11,10%
14	10	145	37					1	110	10	0,00%	1	20	10	0,00%
15					Сумма квадратов разности =										
16	Коэффициент ранговой корреляции		Рху=												

Рис. 55. Общий вид размещения результатов работы на листе *«Спирмен»*

5.3. ТЕОРЕТИЧЕСКИЕ ОСНОВЫ РЕГРЕССИИ

Краткие сведения из теории статистики

Выше были рассмотрены основные аспекты корреляционного анализа, задачей которого является определение силы и направления связи между изучаемыми величинами. Наряду с корреляционным анализом может проводиться и *регрессионный анализ*, который заключается в нахождении уравнения связи зависимой случайной величины *Y* (называемой *результативным признаком, откликом)* с независимыми случайными величинами $X_1,X_2,...,X_m$ (называемыми также *факторами, предикторами или регрессорами)*.

Методы регрессионного анализа позволяют по имеющимся данным предсказывать результаты, т.е. ориентированы на планирование и прогнозирование. Цель регрессионного анализа: адекватно связать выходные (*Y)* — зависимые переменные, с входными (*X)* — независимыми.

Форма связи результативного признака *Y* с факторами $X_1, X_2, ..., X_m$ получила название *уравнения регрессии.* В зависимости от типа выбранного уравнения различают *линейную* и *нелинейную* регрессию.

Количество взаимосвязанных признаков определяет *простую* (*парную)* или *сложную (множественную)* регрессию. Если присутствует исследуемая связь между двумя признаками (результативным (*Y)* и факторным (*X)*), то регрессия называется *простой*, если между тремя и более признаками — *сложной* (*множественной, многофакторной)* регрессией.

В модели *линейной множественной* регрессии предполагается, что значения отклика (Y), принимаемые им на рассматриваемом множестве объектов, связаны со значением предикторов (X_i) на этих объектах с помощью системы линейных уравнений. В упрощенном виде этот процесс можно

представить в виде одного-единственного уравнения регрессии.

Суммарный уровень взаимосвязи (Y) и (X_i) оценивается по величине *коэффициентов множественной корреляции* — ***R*** или *множественной детерминации* — $\boldsymbol{R^2}$**.**

Коэффициент *множественной детерминации* — $\mathbf{R^2}$ является одним из *основных показателей качества регрессии*. Он показывает, с какой степенью точности полученное регрессионное уравнение аппроксимирует (описывает) исходные данные. Он принимает значения от нуля до единицы, чем ближе его значение к единице, тем выше качество регрессии. Если $\mathbf{R^2 > 0{,}95}$, то говорят о высокой точности аппроксимации (модель хорошо описывает явление). Если $\mathbf{0{,}8 < R^2 < 0{,}95}$, то говорят об удовлетворительной аппроксимации (модель в целом адекватна описываемому явлению). Если $\mathbf{R^2 < 0{,}60}$, то принято считать, что точность аппроксимации недостаточна и модель требует улучшения (введения новых независимых переменных, учета нелинейностей).

Коэффициент *множественной корреляции* — **R** также принимает значения в диапазоне от нуля до единицы, чем он ближе к единице, тем выше качество регрессии.

Чем ближе значения этих двух коэффициентов по абсолютным величинам, тем ближе линия регрессии к линейной зависимости между анализируемыми переменными, чем больше разница — тем более вероятна между ними криволинейная зависимость.

Достоверность модели (ее ***общее качество***), т.е. результаты регрессионного анализа оцениваются ***по уровню значимости критерия Фишера p***, который приведен в таблице «*Суммарные результаты*», выдаваемой регрессионным анализом. ***Критерий Фишера p*** должен быть меньше, чем 0,05.

При изучении регрессии следует придерживаться определенной последовательности этапов:

1. Задание аналитической формы уравнения регрессии и определение параметров регрессии.
2. Определение в регрессии степени взаимосвязи результативного признака и факторов, проверка общего качества уравнения регрессии.
3. Проверка статистической значимости каждого коэффициента уравнения регрессии и определение их доверительных интервалов.

Основное содержание выделенных этапов рассмотрим на примере множественной линейной регрессии, реализованной в процедуре «***Регрессия***» надстройки пакет Анализа Microsoft Excel.

Этап 1. Уравнение линейной множественной регрессии имеет вид:

$$\hat{y} = a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_m x_m ,$$

где $\hat{y}$ — теоретические значения результативного признака, полученные путем подстановки соответствующих значений факторных признаков в уравнение регрессии;

$x_1,\ x_2, \ldots, x_m$ — значения факторных признаков;

a_0, a_1,…, a_m — параметры уравнения (коэффициенты регрессии).

Параметры уравнения регрессии определяются с помощью *метода наименьших квадратов*. Сущность данного метода заключается в нахождении параметров модели *(a_i)*, при которых минимизируется сумма квадратов отклонений фактических значений результативного признака от теоретических, полученных по выбранному уравнению регрессии.

Этап 2. Для определения степени взаимосвязи результативного признака *Y* и факторов *X* необходимо знать следующие дисперсии:

- *общую дисперсию* результативного признака Y, отображающую влияние как основных, так и остаточных факторов:

$$\sigma_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n},$$

где $\bar{y}$ — среднее значение результативного признака *Y;*

- *факторную дисперсию* результативного признака *Y,* отображающую влияние только основных факторов:

$$\sigma_\phi^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{n};$$

- *остаточную дисперсию* результативного признака *Y,* отображающую влияние только остаточных факторов:

$$\sigma_o^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-(m+1)}$$

При корреляционной связи результативного признака и факторов выполняется соотношение: $\sigma_\phi^2 < \sigma_y^2$, при этом $\sigma_y^2 = \sigma_\phi^2 + \sigma_o^2$.

Для анализа общего качества уравнения линейной многофакторной регрессии используют обычно *множественный коэффициент детерминации* R^2. Множественный коэффициент детерминации рассчитывается по формуле:

$$R^2 = \frac{\sigma_\phi^2}{\sigma_y^2}$$

и определяет *долю вариации* результативного признака, обусловленную изменением факторных признаков, входящих в многофакторную регрессионную модель.

Так как в большинстве случаев уравнение регрессии приходится строить на основе выборочных данных, то возникает вопрос об адекватности построенного уравнения генеральным данным. Для этого проводится проверка статистической значимости коэффициента детерминации R^2 на основе F-критерия Фишера:

$$F = \frac{R^2}{1-R^2} * \frac{n-m-1}{m},$$

где n — число наблюдений;

m — число факторов в уравнении регрессии.

Примечание. Если в уравнении регрессии свободный член $a_0 = 0$, то числитель $n - m - 1$ следует увеличить на 1, т.е. он будет равен $n - m$.

При значениях $R^2 > 0{,}7$ считается, что вариация результативного признака Y обусловлена в основном влиянием включенных в регрессионную модель факторов X.

Этап 3. Возможна ситуация, когда часть вычисленных коэффициентов регрессии не обладает необходимой степенью значимости. В этом случае такие коэффициенты должны быть исключены из уравнения регрессии. Проверка адекватности построенного уравнения регрессии наряду с проверкой значимости коэффициента детерминации R^2 включает в себя также и проверку значимости каждого коэффициента регрессии для X_i.

Справочная информация по технологии работы

Режим работы «Регрессия» служит для расчета параметров уравнения *линейной* регрессии и проверки его адекватности исследуемому процессу.

В диалоговом окне данного режима (рис. 56) задаются следующие параметры:

1. *Входной интервал Y* — вводится ссылка на ячейки, содержащие данные по результативному признаку. Диапазон должен состоять из одного столбца.
2. *Входной интервал X* — вводится ссылка на ячейки, содержащие факторные признаки. Максимальное число входных диапазонов (столбцов) равно 16.
3. *Метки в первой строке/Метки в первом столбце.*

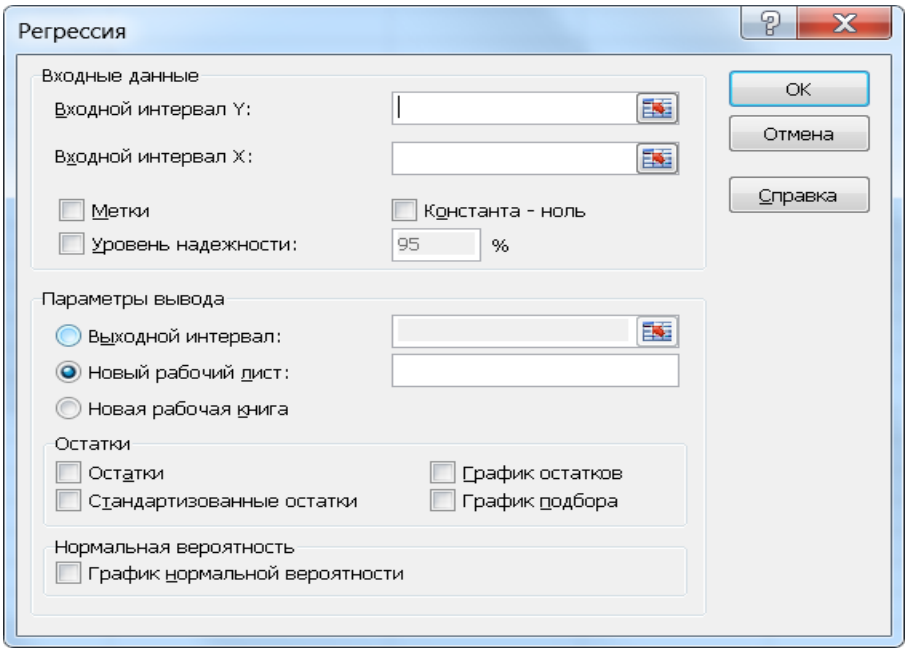


Рис. 56. Окно процедуры «Регрессия»

4. *Уровень надежности* — установите данный флажок в активное состояние. В поле, расположенном напротив флажка введите уровень надежности, если он отличен от уровня 95%, применяемого по умолчанию. Установленный уровень надежности используется для проверки значимости коэффициента детерминации R^2 и коэффициентов регрессии a_i.

5. *Константа-ноль* — установите данный флажок в активное состояние, если требуется, чтобы линия регрессии прошла через начало координат (т.е. $a_0 = 0$ — значение a_0 не является статистически значимым).

6. *Выходной интервал/Новый рабочий лист/Новая рабочая книга.*

7. *Остатки* — установите данный флажок в активное состояние, если требуется включить в выходной диапазон столбец остатков *(см.* столбец *Остатки*).

8. *График остатков* — установите данный флажок в активное состояние, для вывода на Рабочий лист точечных графиков зависимости остатков от факторных признаков x_i.

9. *График подбора* — установите данный флажок в активное состояние, если требуется вывести на Рабочий лист точечные графики зависимости теоретических результативных значений y от факторных признаков x_i.

В результате выполнения процедуры «Регрессия» на Рабочий лист выводится ряд таблиц, данные которых свидетельствуют о качестве полученной математической модели, значениях и значимости коэффициентов линейного уравнения.

Анализ результатов выполнения процедуры «Регрессия»

Работу процедуры и анализ ее результатов рассмотрим на следующем примере.

Постановка задачи 1

Дано: В отделе снабжения больницы имеется информация об изменении стоимости антисептического средства за длительный период времени. В таблице 14 приведены стоимость упаковки антисептика (в руб.) и соответствующий курс доллара (руб./USD).

Таблица 14.

Сведения о стоимости упаковки антисептического средства

Стоимость упаковки	Курс доллара
500	473
700	676
900	901
1200	1126
1500	1427
1600	1577
2000	1877
2500	2200

Требуется:

1. Сопоставляя цену упаковки с изменениями курса доллара за этот же период времени, построить регрессионное уравнение.
2. Вычислить стоимость антисептического средства при курсе доллара 10600 рублей.

Решение задачи

1. Исходные данные введем в таблицу Microsoft Excel и разместим на Рабочем листе в соответствии с рисунком 57. В данном примере зависимая переменная Y — стоимость упаковки, независимая переменная X — курс доллара.

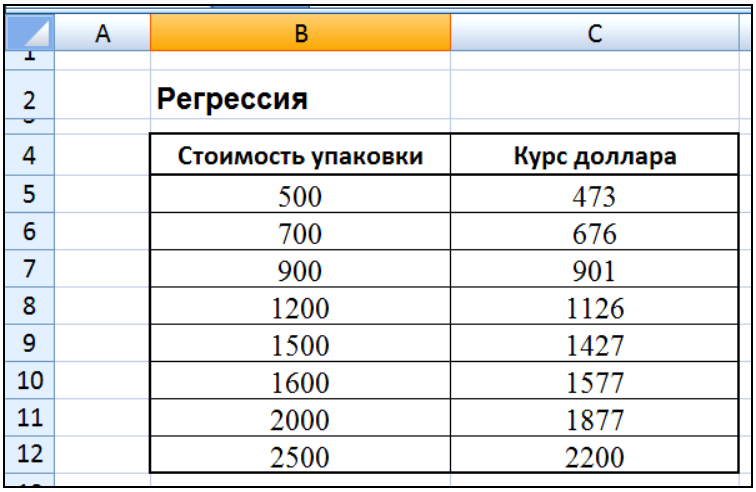


Рис. 57. Вид данных на Рабочем листе Excel

2. Вызовем процедуру «***Регрессия»*** из пакета Анализ данных. Заполним поля окна процедуры в соответствии с рисунком 58.

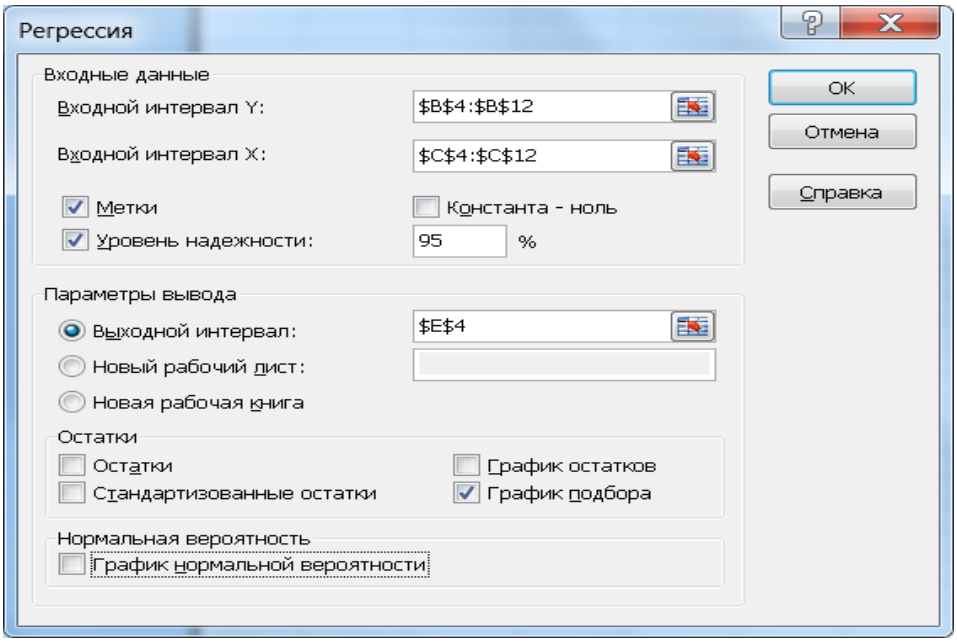
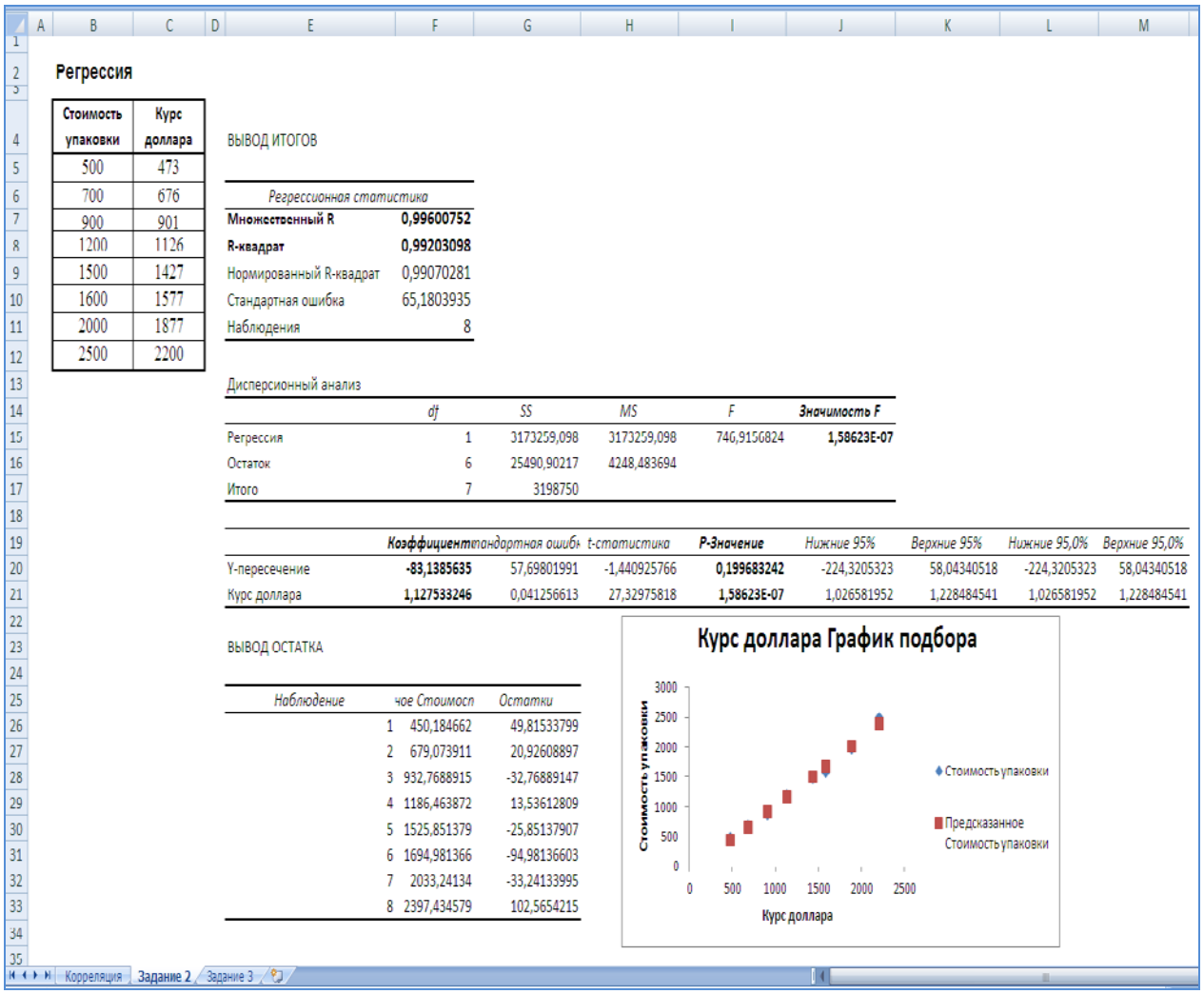


Рис. 58. Вид окна процедура «Регрессия» с заполненными полями

3. Проведем анализ полученных результатов. Их вид на Рабочем листе представлен на рисунке 59.



Регрессия

Стоимость упаковки	Курс доллара
500	473
700	676
900	901
1200	1126
1500	1427
1600	1577
2000	1877
2500	2200

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множественный R	**0,99600752**
R-квадрат	**0,99203098**
Нормированный R-квадрат	0,99070281
Стандартная ошибка	65,1803935
Наблюдения	8

Дисперсионный анализ

	df	*SS*	*MS*	*F*	***Значимость F***
Регрессия	1	3173259,098	3173259,098	746,9156824	**1,58623E-07**
Остаток	6	25490,90217	4248,483694		
Итого	7	3198750			

	Коэффициент	*тандартная ошиб*	*t-статистика*	***P-Значение***	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*
Y-пересечение	**-83,1385635**	57,69801991	-1,440925766	**0,199683242**	-224,3205323	58,04340518	-224,3205323	58,04340518
Курс доллара	**1,127533246**	0,041256613	27,32975818	**1,58623E-07**	1,026581952	1,228484541	1,026581952	1,228484541

ВЫВОД ОСТАТКА

Наблюдение	*чое Стоимосп*	*Остатки*
1	450,184662	49,81533799
2	679,073911	20,92608897
3	932,7688915	-32,76889147
4	1186,463872	13,53612809
5	1525,851379	-25,85137907
6	1694,981366	-94,98136603
7	2033,24134	-33,24133995
8	2397,434579	102,5654215

Рис. 59. Вид Рабочего листа с полученными результатами регрессии

4. Оценим **качество регрессии** по коэффициентам **R** и $\mathbf{R^2}$. Их значения приведены в результатах регрессии в таблице «*Регрессионная статистика*» (рис. 60).

Регрессионная статистика	
Множественный R	**0,99600752**
R-квадрат	**0,99203098**
Нормированный R-квадрат	0,99070281
Стандартная ошибка	65,18039348
Наблюдения	8

Рис. 60. Вид таблицы «Регрессионная статистика»

Значение коэффициентов ***множественной корреляции*** **R = 0, 996 и** ***множественный коэффициент детерминации*** $\mathbf{R^2}$**= 0,992,** что свидетельствует о ***высоком качестве регрессии***. Эти два показателя близки, следовательно, модель очень близка к линейной.

5. Оценим **уровень статистической значимости коэффициента множественной корреляции.**

Он оценивается по значению **F-критерия Фишера.** Значимость F находится в итогах регрессии в таблице *«Дисперсионный анализ»* (рис. 61). Если значение ***Значимость F <0,05***, то коэффициент множественной корреляции статистически значим.

Дисперсионный анализ					
	df	*SS*	*MS*	*F*	***Значимость F***
Регрессия	1	3173259,098	3173259,09	746,91568	**1,58623E-07**
Остаток	6	25490,902	4248,4836		
Итого	7	3198750			

Рис. 61. Вид таблицы *«Дисперсионный анализ»*

В нашем случае **Значимость F - значение p=1,59*10^{-7}** (так следует интерпретировать число 1,58623E-07), что значительно меньше 0,05, следовательно, коэффициент множественной корреляции статистически значим.

6. Рассмотрим **значения коэффициентов регрессионного уравнения** и **их значимость** (рисунке 62, в столбце «Коэффициенты».

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	***P-Значение***	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*
Y-пересечение	**-83,139**	57,698	-1,441	**0,200**	-224,321	58,043	-224,321	58,043
Курс доллара	**1,128**	0,041	27,330	**0,000**	1,027	1,228	1,027	1,228

Рис. 62. Вид таблицы с коэффициентами регрессионного уравнения.

На пересечении строки «Y-пересечение» и столбца «Коэффициенты» находится значение $\mathbf{a_0 = -83{,}139.}$

Значение $\mathbf{a_1}$ находится на пересечении строки «Курс доллара» и столбца «Коэффициенты», $\mathbf{a_1 = 1{,}128.}$

Если значения этих коэффициентов были бы значимы, то уравнение регрессии имело бы вид: **Y= – 83,139 + 1,128 * X.**

Проверим **значимость коэффициентов этого уравнения.** Для этого проанализируем в строках с соответствующими коэффициентами числа в столбце «***p-Значение*».

Для $\mathbf{a_0}$**: p-Значение = 0,2,**что больше 0,05, следовательно коэффициент $\mathbf{a_0}$ является не значимым, его следует отбросить.

Для $\mathbf{a_1}$**: p-Значение = 0,000,** что меньше 0,05, следовательно коэффициент $\mathbf{a_1}$ является значимым. Но в связи с тем, что в данном случае $\mathbf{a_0} = 0$, регрессионное уравнение следует получить заново, повторив вызов процедуры «Регрессия». В окне процедуры *при повторном пересчете* следует установить флажок «***Константа ноль»***. Новое уравнение будет иметь вид:

$$Y = a_1 * X .$$

Получив новое значение $\mathbf{a_1}$ и подставив значение X = 10600, мы сможем вычислить значение Y и ответить на второй вопрос задачи, сколько будет стоить упаковка антисептического средства при новом курсе доллара.

5.4. ПРАКТИЧЕСКОЕ ПРИМЕНЕНИЕ РЕГРЕССИОННОГО АНАЛИЗА ДАННЫХ

ЦЕЛИ ЗАНЯТИЯ

1. Ознакомиться с основными понятиями выявления количественной взаимосвязи признаков — регрессией.
2. Получить навыки применения процедуры «Регрессия» пакета Анализ данных.
3. Научиться записывать по полученным результатам регрессионного анализа уравнение взаимосвязи признаков.
4. Овладеть навыками определения адекватности полученной модели и значимости коэффициентов регрессионного анализа эмпирическим данным.

МЕТОДИКА ВЫПОЛНЕНИЯ РАБОТЫ

Постановка задачи 2

Дано: Ежемесячные данные наблюдений за состоянием погоды и посещаемостью сеансов массажа и гидротерапии размещены на Рабочем листе Excel «*Корреляция*» файла (рис. 63).

	A	B	C	D	E	F	G	H
1		Процедура Корреляция						
2								
3	Число ясных дней	8	14	20	25	20	15	
4	Количество посещений массажа	495	503	380	305	348	465	
5	Количество посещений водного лечения	132	348	643	865	743	541	

Рис. 63. Исходные данные задачи 2

Требуется:

1. Определить, существует ли взаимосвязь между состоянием погоды и посещаемостью массажных и водных процедур, используя функцию КОРРЕЛ(), процедуру «Корреляция» пакета Анализ данных, коэффициент ранговой корреляции Спирмена.
2. Результаты занести в таблицы 15 и 16.
3. По полученным данным сделать вывод о направлении и силе взаимосвязи между количеством солнечных дней — частотой посещения массажа, количеством солнечных дней — частотой посещения гидротерапии.
4. В тетради записать обоснование, какой метод следовало применять в данном случае и почему.

Постановка задачи 3

Дано: В отделе снабжения больницы имеется информация об изменении стоимости антисептического средства за длительный период времени. Данные представлены в таблице 14 и на Рабочем листе *«Задание 2»* файла *Пр.зан.№6-Регрессия.xls*.

Требуется:

1. Построить регрессионное уравнение зависимости стоимости антисептического средства от курса доллара, учитывая, что из изложенного выше известно, коэффициент $\mathbf{a_0 = 0}$.
2. Записать регрессионное уравнение, проанализировать качество модели.
3. Вычислить стоимость антисептического средства при курсе доллара 10600 рублей.
4. Результаты занести в таблицу 16.

Постановка задачи 4

Дано: Данные наблюдений, проводимых в течение 29 месяцев, о содержании в воздухе двуокиси углерода (X_1), степени запыленности (X_2) и уровне заболеваемости органов дыхания (У). Данные представлены Рабочем

листе *«Задание 3»* файла *Пр.зан.№6-Регрессия.xls.*

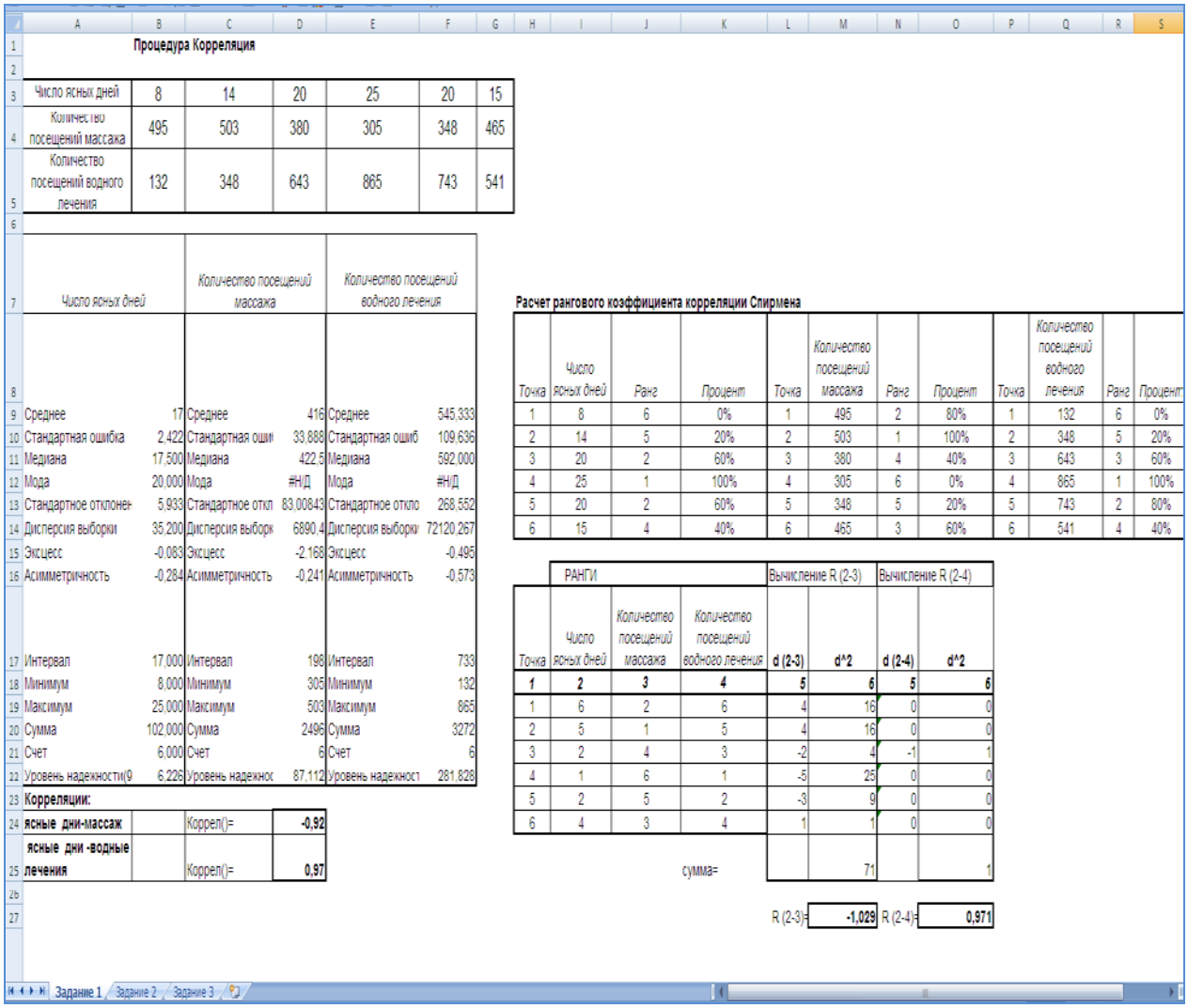
Требуется:

1. Построить регрессионную модель для предсказания изменений уровня заболеваемости органов дыхания (У) в зависимости от содержания в воздухе двуокиси углерода (X_1) и степени запыленности (X_2)**.**
2. Записать регрессионное уравнение, проанализировать качество модели.

ХОД РАБОТЫ

РЕКОМЕНДАЦИИ ПО ВЫПОЛНЕНИЮ ЗАДАНИЙ

1. Скопируйте из папки «Z:\ Материалы для работы\Статистика» в свою папку файл *Пр.зан.№6-Регрессия.xls*, сохраните его с именем *Пр.зан.№6-Иванов А.-24 леч,* подставив свою фамилию.
2. Выполняйте задание на Рабочем листе с именем *«Задание 1»*.
3. Разместите результаты на Рабочем листе *«Задание 1»*, общий вид которого представлен на рисунке 64.



Процедура Корреляция

Число ясных дней	8	14	20	25	20	15
Количество посещений массажа	495	503	380	305	348	465
Количество посещений водного лечения	132	348	643	865	743	541

	Число ясных дней		*Количество посещений массажа*		*Количество посещений водного лечения*
Среднее	17	Среднее	416	Среднее	545,333
Стандартная ошибка	2,422	Стандартная ошиі	33,888	Стандартная ошиб	109,636
Медиана	17,500	Медиана	422,5	Медиана	592,000
Мода	20,000	Мода	#Н/Д	Мода	#Н/Д
Стандартное отклонен	5,933	Стандартное откл	83,00843	Стандартное откло	268,552
Дисперсия выборки	35,200	Дисперсия выборк	6890,4	Дисперсия выборки	72120,267
Эксцесс	-0,083	Эксцесс	-2,168	Эксцесс	-0,495
Асимметричность	-0,284	Асимметричность	-0,241	Асимметричность	-0,573
Интервал	17,000	Интервал	198	Интервал	733
Минимум	8,000	Минимум	305	Минимум	132
Максимум	25,000	Максимум	503	Максимум	865
Сумма	102,000	Сумма	2496	Сумма	3272
Счет	6,000	Счет	6	Счет	6
Уровень надежности(9	6,226	Уровень надежнос	87,112	Уровень надежност	281,828

Корреляции:

ясные дни-массаж		Коррел()=	**-0,92**
ясные дни -водные лечения		Коррел()=	**0,97**

Расчет рангового коэффициента корреляции Спирмена

Точка	*Число ясных дней*	*Ранг*	*Процент*	*Точка*	*Количество посещений массажа*	*Ранг*	*Процент*	*Точка*	*Количество посещений водного лечения*	*Ранг*	*Процент*
1	8	6	0%	1	495	2	80%	1	132	6	0%
2	14	5	20%	2	503	1	100%	2	348	5	20%
3	20	2	60%	3	380	4	40%	3	643	3	60%
4	25	1	100%	4	305	6	0%	4	865	1	100%
5	20	2	60%	5	348	5	20%	5	743	2	80%
6	15	4	40%	6	465	3	60%	6	541	4	40%

	РАНГИ			Вычисление R (2-3)		Вычисление R (2-4)	
Точка	*Число ясных дней*	*Количество посещений массажа*	*Количество посещений водного лечения*	**d (2-3)**	**d^2**	**d (2-4)**	**d^2**
1	**2**	**3**	**4**	**5**	**6**	**5**	**6**
1	6	2	6	4	16	0	0
2	5	1	5	4	16	0	0
3	2	4	3	-2	4	-1	1
4	1	6	1	-5	25	0	0
5	2	5	2	-3	9	0	0
6	4	3	4	1	1	0	0
			сумма=		71		1

R (2-3)= **-1,029** R (2-4)= **0,971**

Задание 1 | Задание 2 | Задание 3

Рис. 64. Вид Рабочего листа с результатами задания 1

4. Вычислите описательные статистики. Результаты запишите в тетрадь в таблицу 15.

5. Определите нормальность частотных распределений исследуемых переменных. Решите, какие методы обработки данных следует применять в данном случае (параметрические, непараметрические) и запишите в таблицу 15.

6. Вычислите коэффициенты корреляции между состоянием погоды и посещением массажных и водных процедур с помощью функции КОРРЕЛ() и процедуры «Корреляция» пакета Анализ данных.

7. При вычислении коэффициентов ранговой корреляции Спирмена используйте алгоритм, представленный в практической работе 5.

8. Выводы о наличии и силе взаимосвязей заданных признаков запишите в тетрадь в таблицу 16.

Таблица 15

Данные для определения нормальности частотных распределений

Признак	Среднее	Медиана	Мода	Вывод о нормальности распределения
Число ясных дней				
Количество посещений массажа				
Количество посещений водного лечения				

Таблица 16.

Связь между признаками	Применяемый метод: (параметрический-непараметрический)	Корреляция		Адек-ватен метод	Вывод о связи
		Пирсона	Спирмена		
Числом **ясных дней** и **количеством посещений массажа**					
Числом **ясных дней** и **количеством посещений водного лечения**					

Решение задачи 3

1. На Рабочем листе *«Задание 2»* проведите регрессионный анализ, использовав процедуру «Регрессия» пакета Анализ данных, выявите зависимость между стоимостью упаковки антисептического средства и курсом доллара, считая известным, что свободный член в выражении (константа - 0) равен нулю.

2. Запишите в тетрадь уравнение регрессии. Обоснуйте значимость важных коэффициентов и качество полученной модели.

3. Вычислите стоимость упаковки антисептического средства при курсе доллара равном 10600 белорусских рублей.

4. Результаты вычисления занесите в таблицу 17.

Таблица 17.

Результаты регрессионного анализа

№ задачи	Р модели/ значимость	R	R^2	Линей-ность	a_0	a_1	a_2	Уравнение: $У=a_0+ a_1x_1+ a_2x_2$
					Pa_0	Pa_1	Pa_2	
3							-	
							-	
4								

Решение задачи 4

1. На Рабочем листе *«Задание 3»* проведите регрессионный анализ, выявив зависимость между частотой заболеваемости органов дыхания (Y) от содержания в воздухе двуокиси углерода (X_1) и степени запыленности (X_2). Ввод параметров процедуры «Регрессия» осуществляйте в соответствии с рисунком 65.

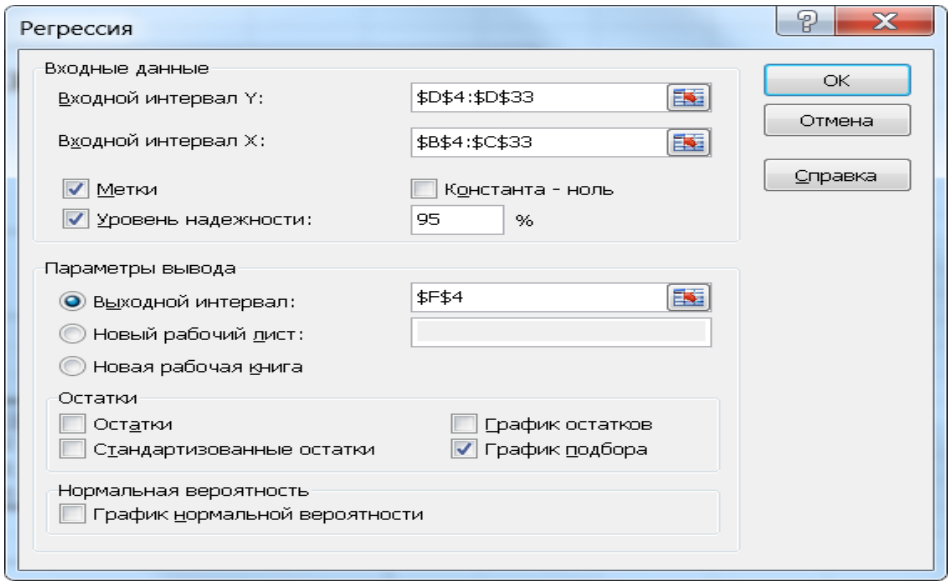
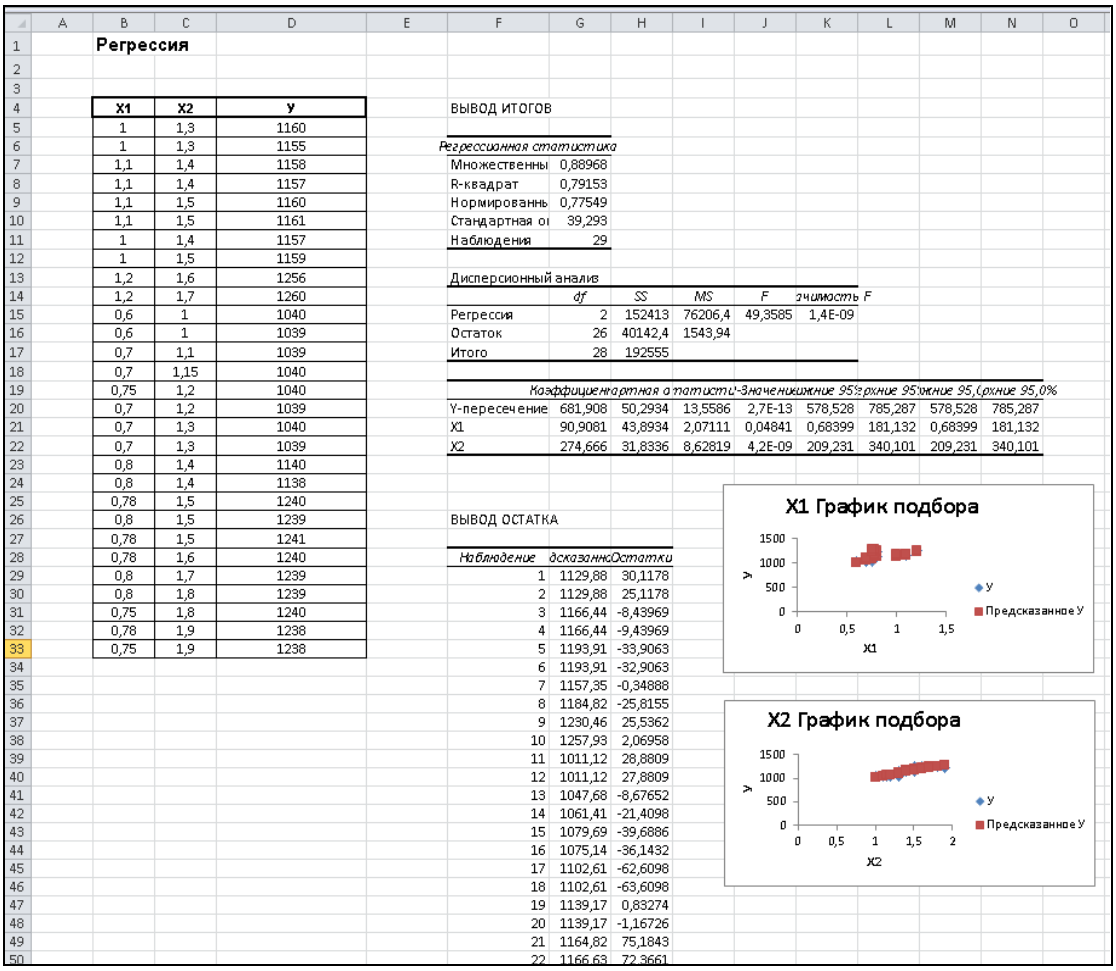


Рис. 65. Окно процедуры «Регрессия» для решения задачи 4

2. Разместите результаты работы в соответствии с рис. 66.



X1	X2	У
1	1,3	1160
1	1,3	1155
1,1	1,4	1158
1,1	1,4	1157
1,1	1,5	1160
1,1	1,5	1161
1	1,4	1157
1	1,5	1159
1,2	1,6	1256
1,2	1,7	1260
0,6	1	1040
0,6	1	1039
0,7	1,1	1039
0,7	1,15	1040
0,75	1,2	1040
0,7	1,2	1039
0,7	1,3	1040
0,7	1,3	1039
0,8	1,4	1140
0,8	1,4	1138
0,78	1,5	1240
0,8	1,5	1239
0,78	1,5	1241
0,78	1,6	1240
0,8	1,7	1239
0,8	1,8	1239
0,75	1,8	1240
0,78	1,9	1238
0,75	1,9	1238

ВЫВОД ИТОГОВ

Регрессионная статистика

Множественны	0,88968
R-квадрат	0,79153
Нормированны	0,77549
Стандартная ои	39,293
Наблюдения	29

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*значимость F*
Регрессия	2	152413	76206,4	49,3585	1,4E-09
Остаток	26	40142,4	1543,94		
Итого	28	192555			

	Коэффициенты	*тартная о*	*татистич*	*Значение*	*нижние 95%*	*рхние 95%*	*нижние 95,0%*	*рхние 95,0%*
Y-пересечение	681,908	50,2934	13,5586	2,7E-13	578,528	785,287	578,528	785,287
X1	90,9081	43,8934	2,07111	0,04841	0,68399	181,132	0,68399	181,132
X2	274,666	31,8336	8,62819	4,2E-09	209,231	340,101	209,231	340,101

ВЫВОД ОСТАТКА

Наблюдение	*дсказанна*	*Остатки*
1	1129,88	30,1178
2	1129,88	25,1178
3	1166,44	-8,43969
4	1166,44	-9,43969
5	1193,91	-33,9063
6	1193,91	-32,9063
7	1157,35	-0,34888
8	1184,82	-25,8155
9	1230,46	25,5362
10	1257,93	2,06958
11	1011,12	28,8809
12	1011,12	27,8809
13	1047,68	-8,67652
14	1061,41	-21,4098
15	1079,69	-39,6886
16	1075,14	-36,1432
17	1102,61	-62,6098
18	1102,61	-63,6098
19	1139,17	0,83274
20	1139,17	-1,16726
21	1164,82	75,1843
22	1166,63	72,3661

Рис. 66. Вид Рабочего листа с результатами решения задачи 4

3. Обоснуйте значимость важных коэффициентов и качество полученной модели.

Анализ модели проводите по следующему алгоритму:

- Оцените **суммарный уровень взаимосвязи** Y и X_i по величине *коэффициентов множественной корреляции — **R*** и *множественной детерминации — **R^2**.*
- Оцените наличие **линейной зависимости** в исходных данных, сравнив значения абсолютных величин *множественной корреляции — **R*** и *множественной детерминации — **R^2**.*
- Проанализируйте **общее качество полученной модели** — ее достоверность по *уровню значимости критерия Фишера(p)*.
- Определите **значения коэффициентов уравнения регрессии:** a_0, a_1, a_2.
- Проанализируйте статистическую значимость этих коэффициентов по соответствующим им значениям p-Значение**.**
- Запишите в тетрадь уравнение регрессии в виде: $Y=a_0+a_1*X_1+a_2*X_2$.
- Результаты занесите в таблицу 17.

4. Определите, какой фактор больше влияет на уровень заболеваемости органов дыхания (У): содержание в воздухе двуокиси углерода (X_1) или запыленность (X_2).

5. Отразите результаты в отчете по практической работе в Вашей тетради в соответствии с алгоритмом анализа модели.

Вопросы для самоконтроля

1. Какая функция Microsoft Excel позволяет вычислить коэффициент корреляции Пирсона?
2. Какая процедура пакета Анализ данных вычисляет параметрический коэффициент корреляции Пирсона?
3. О чем свидетельствует равенство коэффициента корреляции 0?
4. О какой связи между признаками свидетельствует значение коэффициента корреляции равное «– 0,85»?
5. Есть ли программе Microsoft Excel функция и процедура пакета Анализ данных, позволяющие вычислить коэффициент ранговой корреляции Спирмена?
6. Какое максимальное абсолютное значение может принимать коэффициент корреляции?
7. Для каких целей применяется регрессионный анализ данных?
8. Что получают в результате регрессионного анализа?
9. В чем отличие линейного регрессионного анализа от нелинейного?
10. Какую переменную полученного уравнения называют «откликом», а какие – предикторами?
11. Как называют полученные в результате выполнения регрессионного анализа значения R и R^2? О чем они свидетельствуют?

6. ВЫЯВЛЕНИЕ ВЛИЯНИЯ ОТДЕЛЬНЫХ ФАКТОРОВ НА ХОД ПРОФЕССИОНАЛЬНО ЗНАЧИМЫХ МЕДИКО-БИОЛОГИЧЕСКИХ ПРОЦЕССОВ

6.1. ТЕОРЕТИЧЕСКИЕ ОСНОВЫ ДИСПЕРСИОННОГО АНАЛИЗА

Дисперсионный анализ применяется для исследования нескольких выборок на однородность и значимость влияния на процессы отдельных факторов, установления причинно-следственных связей между явлениями. Дисперсионный анализ разработан Р.Фишером. Дисперсионный анализ, который основан на сравнении дисперсий трех и более выборок, применяют, когда следует сравнить их однородность,. В зависимости от числа оказывающих влияние факторов различают *однофакторный* и *многофакторный (двухфакторный* и т. д.) дисперсионный анализ.

Дисперсионный анализ является параметрическим методом: он предполагает, что выборки подчиняются **нормальному закону** частотного распределения. В медицине и биологии это условие очень часто нарушается. В связи с этим был разработан непараметрический аналог дисперсионного анализа для несвязанных выборок — критерий Краскела-Уоллиса, для связанных — критерий Фридмана.

При применении параметрического дисперсионного анализа должны выполняться следующие условия:

1) выборки должны подчиняться *нормальным законам* частотного распределения;
2) выборки должны быть однородны, т.е. иметь *равные дисперсии*;
3) наблюдения независимы и проводятся в одинаковых условиях (выборки *независимы)*, т.е. нельзя предсказать значение какого-либо наблюдения по значению другого.

Как показывает практика, дисперсионный анализ дает корректные результаты даже при нарушении однородности дисперсий, в том случае, если уравнены объемы выборок или отличие их очень незначительно. [12; 13]

Однофакторный дисперсионный анализ

Однофакторный дисперсионный анализ (ОДА) служит для установления влияния отдельного фактора на изменчивость какого-либо признака, значения которого могут быть получены опытным путем в виде случайной величины *Y*. При этом величину *Y* называют *результативным признаком,* а действующий фактор — *фактором А*.

Задачи однофакторного дисперсионного анализа являются самыми простыми, но весьма часто встречаются на практике. Типичный пример — установление зависимости числа респираторных заболеваний у пациентов на участке от удаленности проживания их от центра города. Здесь фактором является удаленность от центра города.

При проведении дисперсионного анализа выдвигаются нулевая и альтернативная гипотезы. *Нулевая гипотеза* ОДА: средние величины у рас-

сматриваемых совокупностей значимо не различаются, *альтернативная* — существуют значимые различия в средних выборок, обусловленные воздействием рассматриваемого фактора. Для анализа значимости полученных результатов используют F-критерий Фишера. *F-критерий Фишера* рассчитывается по следующей формуле:

$$F = \frac{\sigma_{МГ}}{\sigma_{ВГ}}$$

где $\sigma_{МГ}$ — межгрупповая дисперсия (дисперсия групповых средних),

$\sigma_{ВГ}$ — внутригрупповая дисперсия (средняя групповых дисперсий).

Внутригрупповая дисперсия обусловлена случайными величинами, а воздействие фактора проявляется в межгрупповой дисперсии.

Анализируя значение F и *его значимость*, делают вывод о справедливости одной из выдвинутых гипотез.

Однофакторный дисперсионный анализ позволяет по выборочным данным выяснить влияет ли контролируемый фактор на результативный признак.

Если в процессе анализа выявлено влияние фактора ***A*** на результативный признак ***Y***, то можно измерить степень данного влияния с помощью **выборочного коэффициента детерминации**:

$$\rho^2 = \frac{\sigma_{ф}^2}{\sigma_{y}^2}$$

где $\sigma_{ф}^2$ — дисперсия влияющего фактора;

σ_{y}^2 — дисперсия результативного признака.

Этот коэффициент показывает, какая доля выборочной дисперсии σ_{y}^2 объясняется зависимостью результативного признака ***Y*** от влияющего фактора ***A.***

Справочная информация по технологии работы с процедурами, реализующими дисперсионный анализ данных

Процедуры пакета Анализ данных «Двухфакторный дисперсионный анализ без повторений» и «Двухфакторный дисперсионный анализ с повторениями» служат для выяснения на основе выборочных данных факта влияния контролируемых факторов *A* и *B* на результативный признак *Y*.

В диалоговых окнах данных процедур задаются те же параметры, что и в диалоговом окне «*Однофакторный дисперсионный анализ*» (рис. 67) только добавлено поле *Число строк для выборки.* В это поле вводится *число выборок*, приходящихся на каждый уровень одного из факторов. Каждый уровень фактора должен содержать одно и то же количество выборок

(строк/столбцов таблицы).

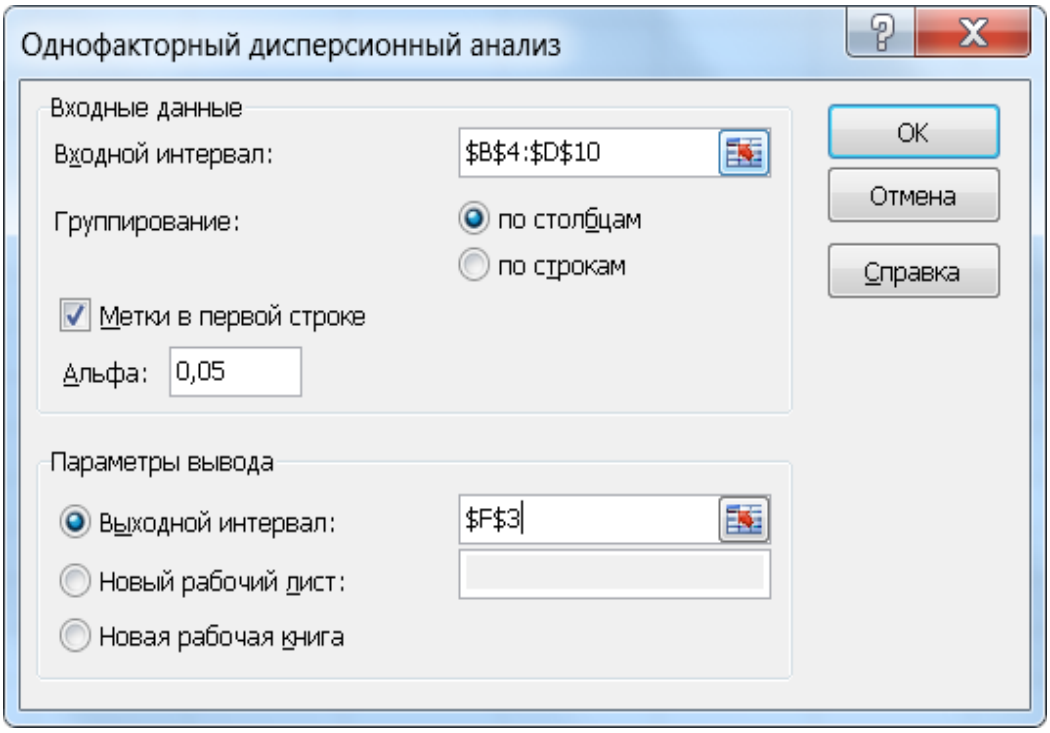


Рис. 67. Вид окна однофакторного дисперсионного анализа

В процедуре «Двухфакторный дисперсионный анализ без повторений» каждому уровню факторов А и В соответствует только одна запись данных, а в процедуре «Двухфакторный дисперсионный анализ с повторениями» каждому уровню факторов соответствует более одной записи данных, причем *число записей для каждого уровня должно быть одинаковым*.

Анализ результатов ОДА

Выполнение однофакторного дисперсионного анализа и анализ его результатов рассмотрим на примере задачи 1.

Постановка задачи 1

Дано: Радиоактивность в крови подопытных животных, подвергавшихся облучению в течение несколько дней, приведена в таблице 18. (размещение данных на Рабочем листе Excel приведено вместе с результатами работы ОДА на рисунке 69).

Таблица 18.

Сведения о радиоактивности в крови животных

День облучения	Радиоактивность в условных единицах			
	1-я группа	2-я группа	3-я группа	4-я группа
1-й	134	130	121	112
2-й	134	130	125	112
3-й	149	157	150	130

Требуется:

Определить влияет ли длительность облучения на изменение радиоактивности в крови животных.

Решение задачи 1

1. Выдвигаем гипотезы:

   H_0: «Отсутствует влияние фактора времени на группу животных»;
   H_1: «Существует влияние фактора времени на группу животных».

2. Вызовем процедуру пакета Анализ данных ***«Однофакторный дисперсионный анализ»***, задав в ее окне значения полей, приведенные на рисунке 68.

✓ **Обратите внимание!** Группировка данных должна производиться по исследуемому фактору. В нашем случае это «День облучения» — группировка осуществляется по строкам.

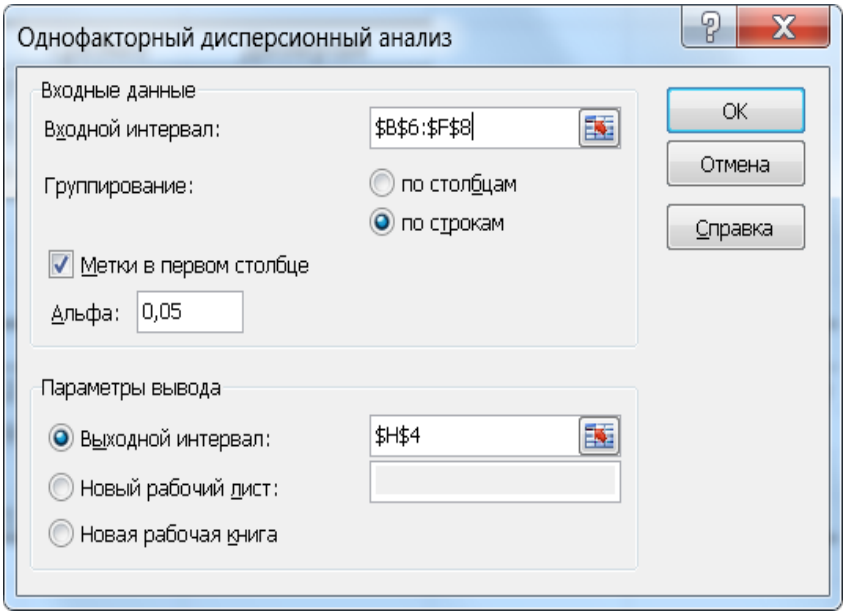


Рис. 68. Окно ОДА с введенными значениями для решения задачи 1

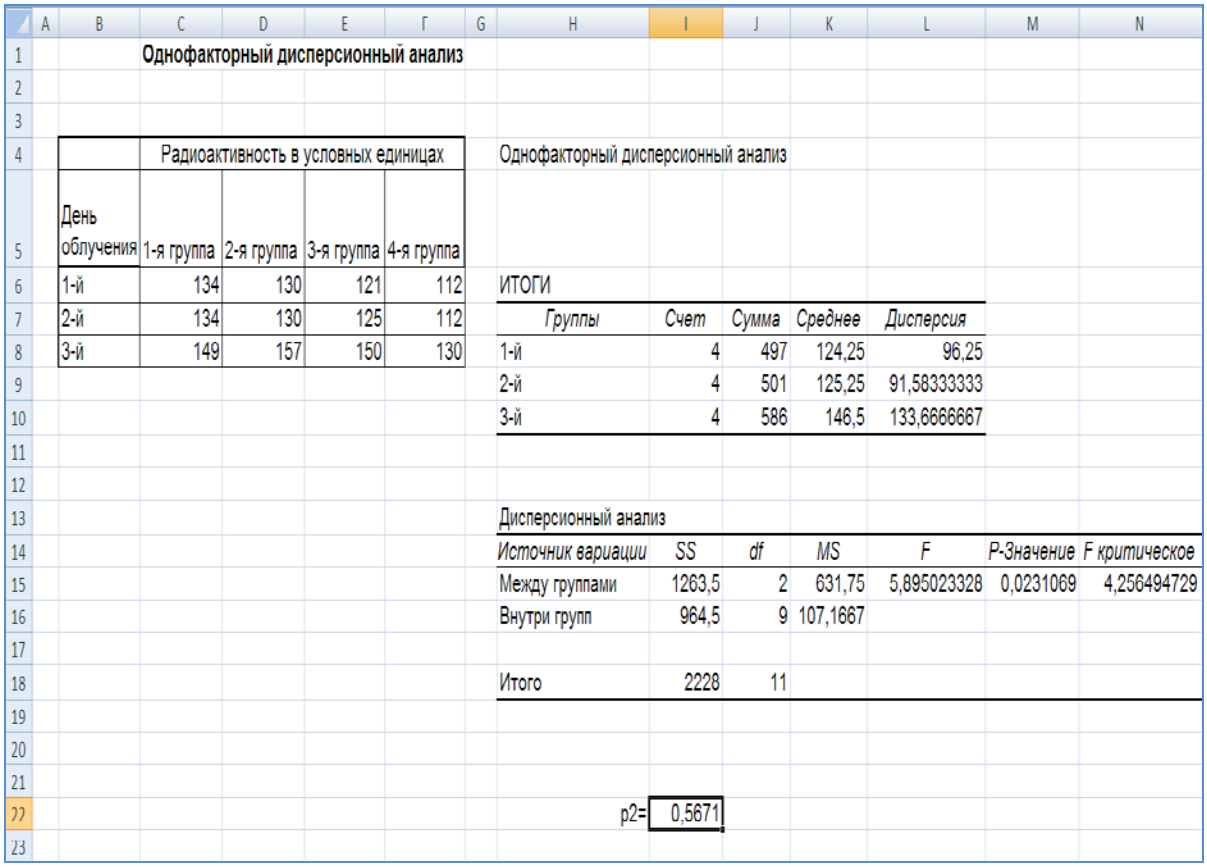
Предварительно необходимо было проверить нормальность частотных распределений в выборках и равенство дисперсий в выборках по факторам. Поскольку количество групп по каждому дню равно, эти проверки можно не выполнять.

Результаты однофакторного дисперсионного анализа представлены на рисунке 69.

3. В результатах ОДА в таблице «*Итоги*» проанализируем значения средних. Значения средних в первый и второй день (1-й день — 124,5; 2-й день — 125,5) практически не отличаются, а в третий день (146,5) значительно выше, чем в первые два дня.

4. Проанализируем *P-Значение* для коэффициента F. Оно равно 0,0231, что < 0,05, следовательно, нулевая гипотеза отвергается — принимается альтернативная, количество дней облучения статистически значимо влияет на

уровень радиации в крови животных. Причем, это влияние практически незначимо в первые два дня (из анализа средних) и значительно отличается в третий день.



Однофакторный дисперсионный анализ

	Радиоактивность в условных единицах			
День облучения	1-я группа	2-я группа	3-я группа	4-я группа
1-й	134	130	121	112
2-й	134	130	125	112
3-й	149	157	150	130

Однофакторный дисперсионный анализ

ИТОГИ

Группы	*Счет*	*Сумма*	*Среднее*	*Дисперсия*
1-й	4	497	124,25	96,25
2-й	4	501	125,25	91,58333333
3-й	4	586	146,5	133,6666667

Дисперсионный анализ

Источник вариации	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*
Между группами	1263,5	2	631,75	5,895023328	0,0231069	4,256494729
Внутри групп	964,5	9	107,1667			
Итого	2228	11				

p2= 0,5671

Рис. 69. Исходные данные и результаты ОДА на рабочем листе Excel

5. Вычислим выборочный коэффициент детерминации. Он равен:

$$\overline{\rho}^2 = \frac{\sigma_\phi^2}{\sigma_y^2}, \qquad \overline{\rho}^2 = \frac{1263{,}5}{2228} = 0{,}567.$$

Это показывает, что 56,7% уровня радиации в крови животных связано с количеством дней облучения.

Вывод: количество радиации в крови животных и количество дней облучения взаимосвязаны.

Многофакторный дисперсионный анализ

Постановка задачи 2

Дано: Радиоактивность в крови подопытных животных, подвергавшихся облучению в течение несколько дней, приведена в таблице 19. (размещение данных на Рабочем листе Excel приведено вместе с результатами выполнения дисперсионного анализа на рисунке 71).

Таблица 19.

Сведения о радиоактивности в крови животных

Радиоактивность в условных единицах				
День облучения	**1-я группа**	**2-я группа**	**3-я группа**	**4-я группа**
1-й	30	28	26	24
1-й	28	30	27	26
1-й	34	32	30	28
1-й	42	40	38	34
2-й	36	38	34	32
2-й	28	30	29	26
2-й	34	32	30	28
2-й	36	30	32	26
3-й	40	38	36	24
3-й	38	36	34	32
3-й	34	45	40	38
3-й	37	38	40	36

Требуется:

Определить влияет ли длительность облучения на изменение радиоактивности в крови животных и зависит ли показатель радиоактивности от группы животных.

Решение задачи 2

1. Выдвигаем гипотезы (для первого фактора *«Дни»*) — А:

   H_0: «Отсутствует влияние фактора времени на группу животных»;
   H_1: «Существует влияние фактора времени на группу животных».

2. Выдвигаем гипотезы (для второго фактора *«Группы»*) — В:

   H_0: «Отсутствует влияние фактора группы на показатель радиоактивности животных»;
   H_1: «Существует влияние фактора группы на показатель радиоактивности животных».

3. Вызовем процедуру ***«Двухфакторный дисперсионный анализ с повторениями»,*** задав в ее окне значения полей, приведенные на рисунке 70.

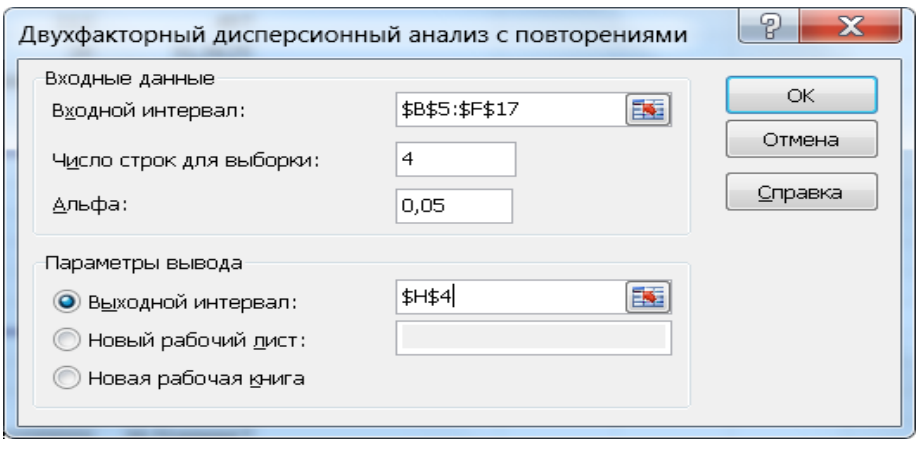
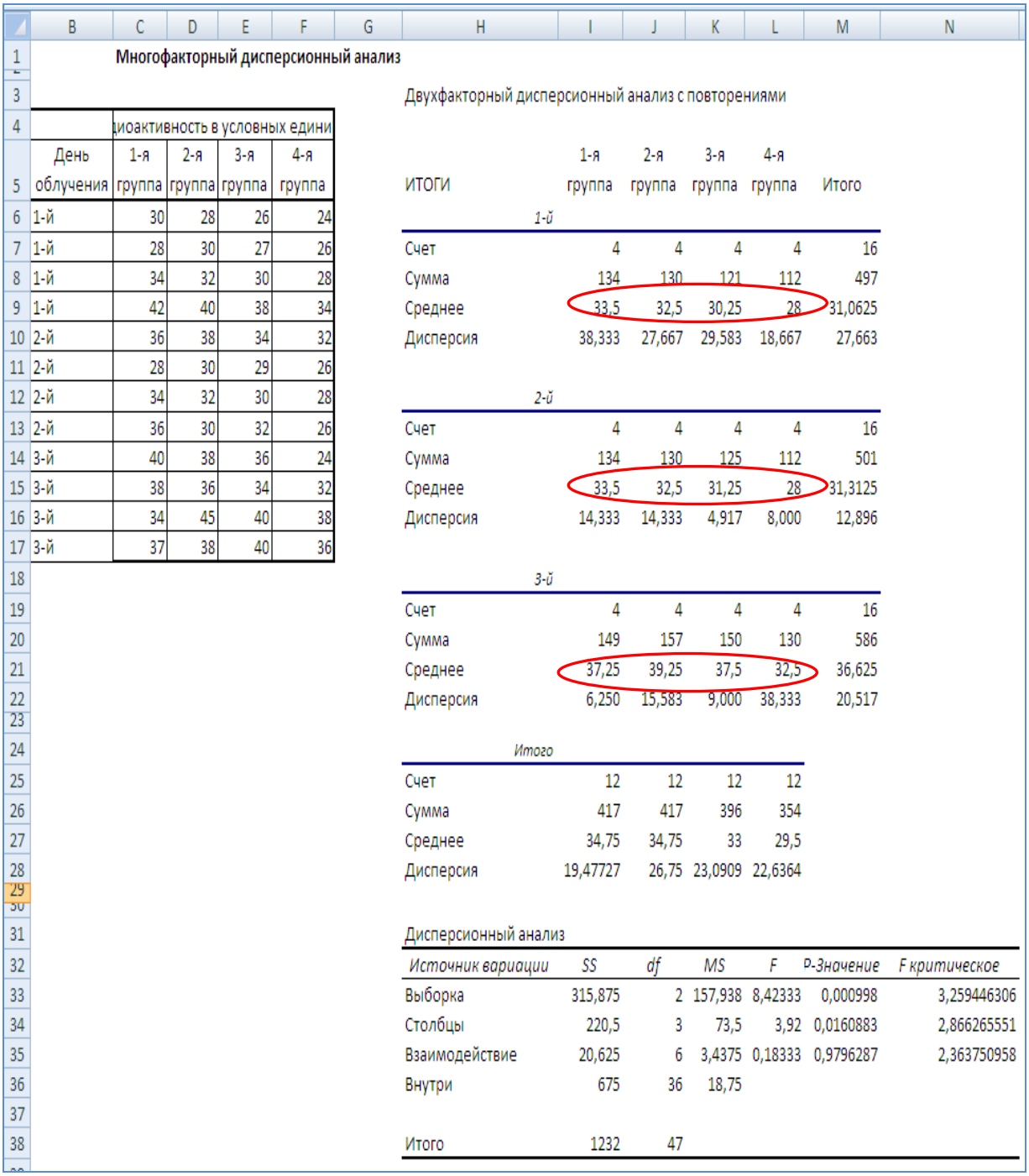


Рис. 70. Вид *окна «Двухфакторного дисперсионного анализа с повторениями»* с введенными значениями для решения задачи 1

Выборки на нормальность частотных распределений и равенство дисперсий не проверяем в связи с тем, что количество повторений в группах одинаково (4) — 4 строки на каждый день облучения. Результаты анализа представлены на рисунке 71. В таблице дисперсионный анализ используются следующие обозначения:

- SS — суммы квадратов отклонений для рассчитываемых параметров;
- MS — средние суммы квадратов отклонений (SS / на степени свободы);
- F — статистика F-критерия;
- P — вероятность нулевой гипотезы.



Многофакторный дисперсионный анализ

День облучения	1-я группа	2-я группа	3-я группа	4-я группа
1-й	30	28	26	24
1-й	28	30	27	26
1-й	34	32	30	28
1-й	42	40	38	34
2-й	36	38	34	32
2-й	28	30	29	26
2-й	34	32	30	28
2-й	36	30	32	26
3-й	40	38	36	24
3-й	38	36	34	32
3-й	34	45	40	38
3-й	37	38	40	36

Двухфакторный дисперсионный анализ с повторениями

ИТОГИ	1-я группа	2-я группа	3-я группа	4-я группа	Итого
1-й					
Счет	4	4	4	4	16
Сумма	134	130	121	112	497
Среднее	33,5	32,5	30,25	28	31,0625
Дисперсия	38,333	27,667	29,583	18,667	27,663
2-й					
Счет	4	4	4	4	16
Сумма	134	130	125	112	501
Среднее	33,5	32,5	31,25	28	31,3125
Дисперсия	14,333	14,333	4,917	8,000	12,896
3-й					
Счет	4	4	4	4	16
Сумма	149	157	150	130	586
Среднее	37,25	39,25	37,5	32,5	36,625
Дисперсия	6,250	15,583	9,000	38,333	20,517
Итого					
Счет	12	12	12	12	
Сумма	417	417	396	354	
Среднее	34,75	34,75	33	29,5	
Дисперсия	19,47727	26,75	23,0909	22,6364	

Дисперсионный анализ

Источник вариации	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*
Выборка	315,875	2	157,938	8,42333	0,000998	3,259446306
Столбцы	220,5	3	73,5	3,92	0,0160883	2,866265551
Взаимодействие	20,625	6	3,4375	0,18333	0,9796287	2,363750958
Внутри	675	36	18,75			
Итого	1232	47				

Рис. 71. Результаты двухфакторного дисперсионного анализа с повторениями

3. Анализируем средние по дням облучения и отдельным группам в таблице «*Итоги*». Видим, что средние в первый и второй день облучения различаются незначительно, но сильно отличаются в третий день.

4. Анализируем таблицу «*Дисперсионный анализ*». В укрупненном виде она приведена на рисунке 72.

Дисперсионный анализ						
Источник вариации	*SS*	*df*	*MS*	*F*	***P-Значение***	*F критическое*
Выборка	315,875	2	157,9375	8,423333333	0,000997995	3,259446306
Столбцы	220,5	3	73,5	3,92	0,01608828	2,866265551
Взаимодействие	20,625	6	3,4375	0,183333333	0,979628738	2,363750958
Внутри	675	36	18,75			
Итого	1232	47				

Рис. 72. Укрупненный вид таблицы «*Дисперсионный анализ*»

P-значение в строке «*Выборка*» показывает значимость влияния на результаты фактора *День облучения* (первый столбец в исходных данных) — фактор ***A***. *P-значение* = 0,000998 < 0,05, следовательно, этот фактор статистически значимо влияет на результат.

P-значение в строке «*Столбцы*» (группы) показывает значимость влияния на результаты фактора *Группа. P-значение* = 0,016 < 0,05, следовательно, и этот фактор (***B***) статистически значимо влияет на результат.

P-значение в строке «*Взаимодействие*» показывает статистическую значимость влияния на результаты взаимодействия этих факторов. *P-значение* = 0,98 > 0,05, следовательно, взаимное влияние сочетания этих факторов статистически значимо не влияет на результат.

Вычислим выборочные коэффициенты детерминации для двух факторов.

Для фактора ***A:*** $\rho^2 = \frac{315{,}875}{1232} = 0{,}256$. Выборочный коэффициент детерминации показывает, что 25,6% изменчивости количества радиации в крови животных связано с влиянием количества дней облучения.

Для фактора ***B:*** $\rho^2 = \frac{220{,}5}{1232} = 0{,}179$. Выборочный коэффициент детерминации показывает, что 17,9% изменчивости количества радиации в крови животных связано с группой животных.

Вывод: количество радиации в крови животных и количество дней облучения взаимосвязаны, количество радиации в крови животных связано с группой животных.

6.2. ПРАКТИЧЕСКОЕ ПРИМЕНЕНИЕ ДИСПЕРСИОННОГО АНАЛИЗА ДАННЫХ

ЦЕЛИ ЗАНЯТИЯ

1. Ознакомиться с основными понятиями дисперсионного анализа данных.
2. Получить навыки применения процедур «Однофакторный дисперсионный анализ», «Двухфакторный дисперсионный анализ без повторений», «Двухфакторный дисперсионный анализ с повторениями» пакета Анализ данных.
3. Научиться делать выводы по полученным результатам дисперсионного анализа о статистической значимости влияния на исследуемый признак анализируемых факторов.
4. Овладеть навыками определения адекватности полученной модели.

МЕТОДИКА ВЫПОЛНЕНИЯ РАБОТЫ

Постановка задачи 3

Дано: Ежемесячные данные наблюдений за состоянием погоды и посещаемостью сеансов массажа и гидротерапии размещены на Рабочем листе Excel *«Задание 3»* файла и представлены на рисунке 73.

	A	B	C	D	E	F	G
1	**Процедура Корреляция**						
2							
3	Число ясных дней	8	14	20	25	20	15
4	Количество посещений массажа	495	503	380	305	348	465
5	Количество посещений водного лечения	132	348	643	865	743	541

Рис. 73. Исходные данные задачи 3

Требуется:

1. В зависимости от заданного преподавателем варианта выполнить одно из двух заданий:

I вариант. Определить с помощью регрессионного анализа наличие значимой количественной взаимосвязи между состоянием погоды и посещаемостью массажных процедур.

II вариант. Определить, с помощью регрессионного анализа наличие значимой количественной взаимосвязи между состоянием погоды и посещаемостью водных процедур.

2. Вычислить коэффициент выборочной детерминации, определяемый этим фактором.

3. Записать, если возможно, регрессионное уравнение, характеризующее выявленную взаимосвязь между количеством ясных дней и количеством посещений массажа (или количеством ясных дней и количеством посещений водного лечения).

Постановка задачи 4

***Дано*:** Введены 3 уровня расстояний от центра города: 1) до 3 км, 2) от 3 до 5 км и 3) свыше 5 км. Данные представлены в таблице 20 и на Рабочем листе *«Задание 4»* файла *Пр.зан.№7- дисперсионный анализ.xls*.

Таблица 20

Удаленность от центра		
до 3 км	**3-5 км**	**свыше 5 км**
92	90	87
98	86	79
89	84	74
97	91	85
90	83	73
94	82	77

Требуется:

1. Выявить значимость влияния расстояния проживания от центра города на частоту респираторных заболеваний помощью *Однофакторного дисперсионного анализа*.
2. Сформулировать и записать в тетрадь аргументированный вывод о влиянии на частоту респираторных заболеваний расстояния проживания пациента от центра города.
3. Вычислить коэффициенты выборочной детерминации, определяемые этими факторами.

Постановка задачи 5

Дано: Для изготовления лекарственного средства в качестве сырья используется выращенная биомасса. В таблице 21 приведены данные по относительному росту (**Р**) 5 образцов биомассы в различных средах: лак, акр, акфа, за разное количество дней: 10, 15, 20, 30.

Требуется:

1. Провести двухфакторный дисперсионный анализ для выявления влияния оптимального сочетания среды и времени выращивания биомассы с целью получения максимального ее количества $Р_{отн}$.
2. Записать в тетрадь аргументированный вывод о значимости влияния среды, времени выращивания биомассы и их сочетания на относительный рост сырья.
3. Вычислить коэффициенты выборочной детерминации, определяемые этими факторами.

Таблица 21

Данные об относительном приросте биомассы в различных средах

Начало таблицы

Среда	Дни	$P_{отн}$
лак	10	0,85
лак	10	0,857143
лак	10	0,85
лак	10	0,894737
лак	10	0,882353
лак	15	0,789474
лак	15	0,8
лак	15	0,9375
лак	15	0,833333
лак	15	0,888889
лак	20	0,545455
лак	20	0,625
лак	20	0,521739
лак	20	0,47619
лак	20	0,6
лак	30	0,454545
лак	30	0,444444
лак	30	0,555556
лак	30	0,571429
лак	30	0,5

Продолжение таблицы

Среда	Дни	$P_{отн}$
акр	10	0,888889
акр	10	0,904762
акр	10	0,777778
акр	10	0,736842
акр	10	0,904762
акр	15	0,685714
акр	15	0,741935
акр	15	0,857143
акр	15	0,785714
акр	15	0,727273
акр	20	0,5
акр	20	0,55
акр	20	0,5
акр	20	0,52381
акр	20	0,541667
акр	30	0,45
акр	30	0,380952
акр	30	0,5
акр	30	0,666667
акр	30	0,409091

Продолжение таблицы

Среда	Дни	$P_{отн}$
акфа	10	0,538462
акфа	10	0,428571
акфа	10	0,352941
акфа	10	0,538462
акфа	10	0,181818
акфа	15	0,666667
акфа	15	0,421053
акфа	15	0,285714
акфа	15	0,470588
акфа	15	0,529412
акфа	20	0,423077
акфа	20	0,4
акфа	20	0,454545
акфа	20	0,55
акфа	20	0,416667
акфа	30	0,5
акфа	30	0,36
акфа	30	0,541667
акфа	30	0,5
акфа	30	0,611111

ХОД РАБОТЫ

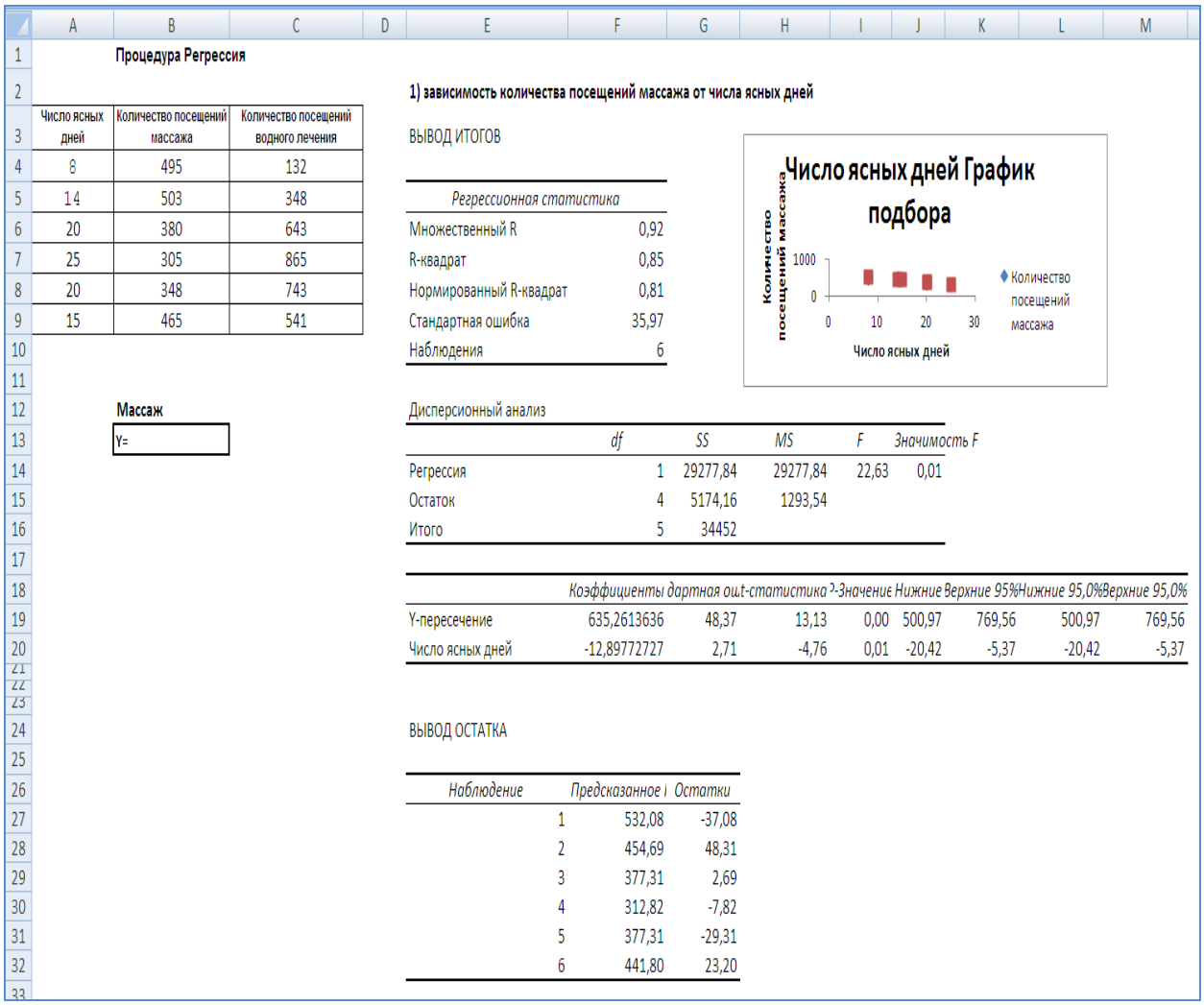
Рекомендации по решению задачи 3

1. Скопируйте из папки «Z:\ Материалы для работы\Статистика» в свою папку файл *Пр.зан.№7-дисперсионный анализ.xls*, сохраните его с именем *Пр.зан.№7- Иванов А. — 24 леч.xls*, подставив свою фамилию.

2. Выполняйте задание на Рабочем листе с именем *«Задание 3»*. Точность вычислений — до трех знаков после запятой.

3. Разместите результаты на Рабочем листе «*Задание 3*», общий вид которого представлен на рисунке 74

4. Выводы в соответствии с полученными результатами в задании 3 запишите в тетрадь.



Процедура Регрессия

Число ясных дней	Количество посещений массажа	Количество посещений водного лечения
8	495	132
14	503	348
20	380	643
25	305	865
20	348	743
15	465	541

Массаж

Y=

1) зависимость количества посещений массажа от числа ясных дней

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множественный R	0,92
R-квадрат	0,85
Нормированный R-квадрат	0,81
Стандартная ошибка	35,97
Наблюдения	6

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	1	29277,84	29277,84	22,63	0,01
Остаток	4	5174,16	1293,54		
Итого	5	34452			

	Коэффициенты	*дартная ош*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95,0%*	*Нижние 95,0%*	*Верхние 95,0%*
Y-пересечение	635,2613636	48,37	13,13	0,00	500,97	769,56	500,97	769,56
Число ясных дней	-12,89772727	2,71	-4,76	0,01	-20,42	-5,37	-20,42	-5,37

ВЫВОД ОСТАТКА

Наблюдение	*Предсказанное i*	*Остатки*
1	532,08	-37,08
2	454,69	48,31
3	377,31	2,69
4	312,82	-7,82
5	377,31	-29,31
6	441,80	23,20

Рис. 74. Вид Рабочего листа с результатами задания 3

Решение задачи 4

1. На Рабочем листе *«Задание 4»* проведите однофакторный дисперсионный анализ, использовав соответствующую процедуру пакета Анализ данных.

2. Проведите анализ полученных результатов: обоснуйте влияние на частоту респираторных заболеваний исследуемых факторов и их сочетания.

3. Запишите в тетрадь аргументированные выводы по результатам дисперсионного анализа.

4. Вычислите коэффициенты выборочной детерминации, определяемые этими факторами, используя формулу:

$$\overline{\rho}^2 = \frac{\sigma_{\phi}^2}{\sigma_y^2}.$$

5. Результаты дисперсионного анализа разместите на Рабочем листе в соответствии с рисунком 75.

	Удаленность от центра		
	до 3 км	3-5 км	свыше 5 км
	92	90	87
	98	86	79
	89	84	74
	97	91	85
	90	83	73
	94	82	77

Однофакторный дисперсионный анализ

ИТОГИ

Группы	*Счет*	*Сумма*	*Среднее*	*Дисперсия*
до 3 км	6	560	93,333	13,467
3-5 км	6	516	86,000	14,000
свыше 5 км	6	475	79,167	32,967

Дисперсионный анализ

Источник вариации	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*
Между группами	602,333	2,000	301,167	14,950	0,0003	3,682
Внутри групп	302,167	15,000	20,144			
Итого	904,5	17				

Рис. 75. Вид рабочего листа с результатами выполнения задачи 4

Решение задачи 5

1. На Рабочем листе *«Задание 5»* проведите дисперсионный анализ, выявив зависимость между средой, временем выращивания и относительным ростом биомассы. Задайте параметры выполнения процедуры в соответствии с рисунком 76, учитывая, что число образцов равно 5.

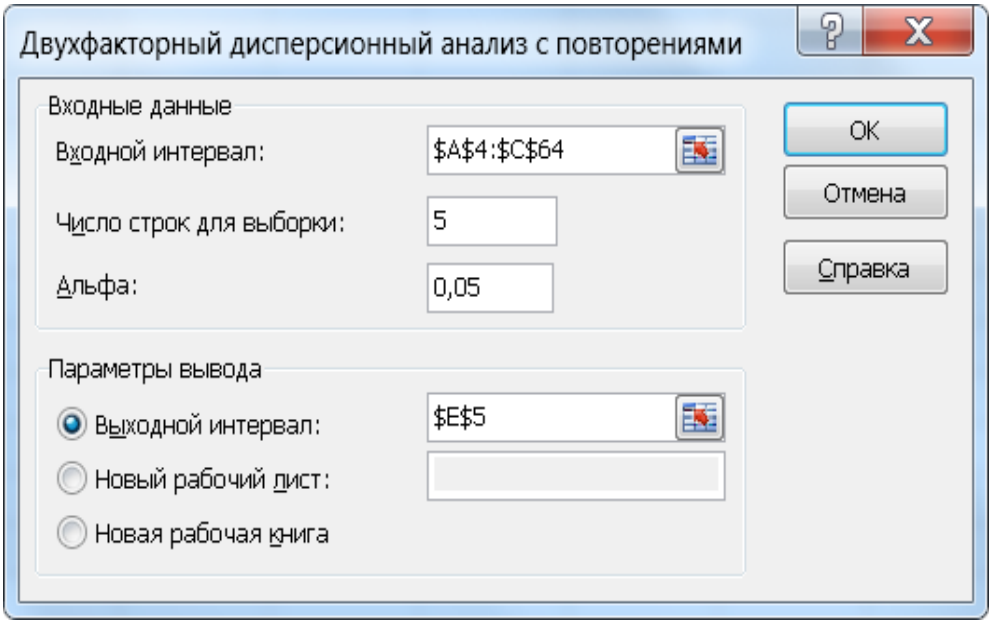


Рис. 76. Вид *окна « Двухфакторного дисперсионного анализа с повторениями»* с введенными значениями для решения задачи 5

2) Проанализируйте в полученной таблице *«Дисперсионный анализ»* значимосоть влияния среды (в столбце *«Источник вариации»* ***P-значение*** в строке «*Выборка»*). Вычислите *выборочный коэффициент детерминации* для этого фактора (*дни*).

3) Проанализируйте в полученной таблице *«Дисперсионный анализ»* значимость влияния дней (в столбце *«Источник вариации»* ***P-значение*** в строке «*Столбцы»*). Вычислите *выборочный коэффициент детерминации* для этого фактора (*среда*).

4) Какой из указанных факторов больше влияет на изменчивость относительного роста биомассы?

5) Сравнивая в результатах дисперсионного анализа среднее значение роста биомассы ($P_{отн}$) за разное количество дней в разных средах, определите среду и количество дней ее выращивания, когда это значение ($P_{отн}$) будет максимально.

Выделите это значение красным цветом в результатах работы.

Запишите в отчет:

«Лучшие значения факторов для выращивания биомассы:»

Среда:_____, Дни:_____; Среднее значение $P_{отн}$:_____

Вопросы для самоконтроля

1. В каких случаях применяется дисперсионный анализ данных?
2. Чем отличается дисперсионный анализ от регрессионного?
3. Какие виды дисперсионного анализа вы знаете?
4. Дисперсионный анализ является параметрическим или не параметрическим методом статистическим методом обработки данных?
5. В каких случаях требование нормальности частотных распределений данных является критичным для выполнения дисперсионного анализа?
6. Какие параметры результатов проведения дисперсионного анализа свидетельствуют о значимости влияния анализируемых факторов на ход исследуемого процесса?
7. Что называется выборочным коэффициентом детерминации? О чем он свидетельствует и как вычисляется?
8. Какой параметр результатов выполнения двухфакторного дисперсионного анализа свидетельствует о значимости влияния двух факторов на ход исследуемого процесса?

ВЫПОЛНЕНИЕ КОМПЛЕКСНОГО ЗАДАНИЯ ИНДИВИДУАЛЬНОЙ РАБОТЫ

ЦЕЛИ ЗАНЯТИЯ

1. Выявить уровень усвоения материала по основам статистического анализа.
2. Сформировать мотивацию к его практическому применению в учебной и профессиональной деятельности.
3. Закрепить полученные при изучении основ статистики навыки.

ЗАДАНИЯ ДЛЯ ВЫПОЛНЕНИЯ

I. Выявление достоверности различий

1. В соответствии с номером варианта работы, полученного у преподавателя, ознакомьтесь с предложенными Вам исходными данными из ***вариантов заданий для выявления достоверности различий*** (стр. 114).
2. Введите данные в соответствии со своим вариантом на Рабочий лист 1 электронной таблицы Microsoft Excel.
3. Выполните ***описательную статистику*** средствами процедуры Анализ данных.
4. На основании полученных данных, используя приблизительные критерии сделайте выводы ***о нормальности*** частотных распределений контрольной и исследуемых групп.
5. Вычислите ***границы доверительных интервалов*** для средних в двух группах.
6. Предположив, что выборки имеют нормальный закон частотного распределения, на основании вычисленных границ доверительных интервалов сделайте предварительный вывод о статистической достоверности различий изучаемого признака в двух группах, при необходимости об эффективности новой методики лечения (лекарственного средства). Вывод запишите в тетрадь.
7. Постройте с помощью процедуры «***Гистограмма»*** пакета Анализ данных гистограммы частотных распределений в двух группах. Сравните, соответствует ли вид полученных гистограмм нормальным законам частотного распределения. Вывод запишите в тетрадь.
8. По результатам описательной статистики определите, в какую сторону относительно нормального наклонены исследуемые частотные распределения выборок? Как отличается пологость вершин частотных распределений групп от нормального распределения?
9. В соответствии с предыдущим анализом запишите в тетрадь, какие методы выявления достоверности различий следует использовать (параметрические или непараметрические).
10. Используя соответствующие встроенные функции Microsoft Excel,

с помощью **основного метода** (критерий Стьюдента или χ^2 Пирсона) вычислите и аргументируйте наличие достоверности различий, и если необходимо об эффективности инноваций.

11. **Альтернативным методом** (параметрическим или непараметрическим), используя соответствующую встроенную функцию Microsoft Excel, проверьте наличие достоверности различий. Запишите в тетрадь вывод о соответствии или различии в полученных разными методами результатах.

12. Проверьте полученные результаты соответствующей **процедурой** пакета Анализ данных, **реализующей критерий Стьюдента**.

✓ **Обратите внимание!** Что, при применении метода χ^2 Пирсона, следует в ряде случаев от признака переходить к частотам.

II. Корреляционный анализ

Постановка задачи

Дано: Возраст и систолическое давление обследованной группы женщин (таблица 22).

Требуется:

Определить зависимость между возрастом женщин и систолическим давлением.

Введите исходные данные в соответствии с таблицей 22 и выполните анализ на Рабочем листе 2.

Таблица 22.

Возраст и систолическое давление у группы женщин

Возраст	Давление
71	173
33	118
31	125
55	155
63	153
49	160
58	148
38	142
36	110
64	142
45	128
68	160
42	136
76	150
34	121
75	166
78	154
62	135
68	146
46	127

Требуется:

1. Определить какие методы следует использовать для обработки данных (параметрические, непараметрические).
2. Вычислить корреляцию основным методом параметрическим или непараметрическим (Пирсона или Спирмена).
3. Вычислить корреляцию альтернативным методом.
4. Сделать вывод о наличии и величине силы связи между возрастом и систолическим давлением.

III. Регрессионный анализ

Постановка задачи

Дано: Данные таблицы 22. (Скопируйте таблицу с Рабочего листа 2 на Рабочий лист 3).

Требуется:

1. Найти регрессионную зависимость между возрастом (переменная X) и систолическим давлением (переменная У) в виде $У = a_0 + a_1 * X_1$.
2. Определить адекватность и значимость модели.
3. Указать значимость коэффициентов.

ВАРИАНТЫ ЗАДАНИЙ ДЛЯ ВЫЯВЛЕНИЯ ДОСТОВЕРНОСТИ РАЗЛИЧИЙ

I вариант

Постановка задачи

Дано: Данные по заболеваемости гриппом детей в школе за 2 года приведены в таблице. В 2000 г. не проводилась вакцинация от гриппа, в 2001 г. — проводилась. Общее количество исследованных детей по каждому году в классах одинаково по 240 человек.

Классы	**Количество заболевших**	
	2000 г.	***2001 г.***
5	10	5
6	8	9
7	12	14
8	15	10
9	8	12
10	14	6
11	18	20
12	11	17

Требуется:

Обосновать выводы об эффективности вакцины.

II вариант

Постановка задачи

Дано: Две группы пациентов с тахикардией. Одна из них (контрольная) получала традиционное лечение, другая (исследуемая) получала лечение по новой методике. В таблице приведены частоты сердечных сокращений (ЧСС) для каждой группы (ударов в минуту). (Норма — 60-80 ударов в минуту).

Частота сердечных сокращений в группах	
контрольная	***исследуемая***
162	135
156	126
144	115
132	140
125	121
151	112
180	130
110	170

Требуется:

Определить эффективность новой методики.

III вариант

Постановка задачи

Дано: Количество заболевших гриппом в старших классах школы №171 среди учащихся, которым проводилась вакцинация от гриппа, и не проводилась. В классах с вакцинацией 240 учащихся, без вакцинации 200 учащихся.

Классы	**Количество заболевших среди групп детей, у которых вакцинация**	
	проводилась	***не проводилась***
5	10	5
6	8	9
7	12	14
8	15	10
9	8	12
10	14	6
11	18	20
12	11	17

Требуется:

Обосновать выводы об эффективности вакцины.

IV вариант

Постановка задачи

Дано: Результаты теста определения эмоциональной составляющей речи в условиях маскировки шумом, полученные в ходе исследования возрастных изменений слуховой функции у детей.

Количество правильных ответов	
Без шума	***6 дБ***
78	61
95	93
80	61
85	73
80	75
90	68
83	70
91	84
89	81

Требуется:

Обосновать выводы о достоверности ухудшения у детей слуха при уровне шума в 6 дБ. В каждой группе таблицы приведены данные для одного испытуемого.

V вариант

Постановка задачи

Дано: Активность холинэстеразы крови у мужчин и женщин (добровольцев) через 2 часа после однократного введения ацетофоса в дозе 2 мг/кг веса (в процентах к исходному фону).

% к исходному фону	
Мужчины	***Женщины***
84,5	51,5
85,6	75,7
92,3	57,6
69,2	68,4
84,1	60,4
70,5	80,1
71,4	79,5

Требуется:

Обосновать выводы о достоверности или недостоверности различия реакции женского и мужского организмов на введенный препарат.

VI вариант

Постановка задачи

Дано: Длительность амбулаторного лечения заболевших гриппом на двух участках с одинаковым количеством пациентов. На первом участке применялись традиционные лекарственные средства. На втором участке применялось новое антибактериальное средство.

Начало таблицы

Участок 1	Участок 2
7	8
3	2
8	7
11	3
9	4
8	8
4	12
7	8
10	9
5	3
11	7

Продолжение таблицы

Участок 1	Участок 2
6	8
7	4
8	5
10	7
8	8
9	5
10	7
4	
9	
11	
7	

Требуется:

Обосновать выводы об эффективности нового лекарственного средства.

VII вариант

Постановка задачи

Дано: Зарплата за месяц медицинского персонала и медработников по конкретной поликлинике.

Персонал	Медработники
210	320
2100	300
200	250
200	200
200	190
190	180
250	400
180	380
220	350
280	
380	

Требуется:

Определить, есть ли статистически значимые различия в заработных платах этих категорий работников. У какой группы она выше?

VIII вариант

Постановка задачи

Дано: Частота заболеваемости органов дыхания на 500 пациентов за одинаковый период времени при различных расстояниях проживания пациентов от центра города.

до 3 км	свыше 5 км
92	87
98	79
89	74
97	85
90	73
94	77
85	90
84	89

Требуется:

Определить, есть ли статистически значимые различия в заболеваемости органов дыхания пациентов при различных расстояниях их проживания от центра города.

IX вариант

Постановка задачи

Дано: Зарплаты врачей нескольких специальностей и медсестер по поликлинике за месяц.

Врачи	Медсестры
450	260
400	210
370	250
300	200
250	300
260	290
230	180
	380

Требуется:

Определить, справедливо ли утверждение, что зарплата врачей в поликлинике статистически значимо больше зарплаты медсестер.

X вариант

Постановка задачи

Дано: Относительный прирост ($P_{отн}$) семи образцов биомассы за разное количество дней.

№ образца	$P_{отн}$	
	10 дней	20 дней
1	0,85	0,789
2	0,857	0,8
3	0,85	0,937
4	0,894	0,833
5	0,882	0,888
6	0,77	0,82
7	0,91	0,89

Требуется:

Определить отличается ли статистически значимо относительный прирост биомассы ($P_{отн}$) за 10 и 20 дней. Справедливо ли утверждение, что за 20 дней он больше?

XI вариант

Постановка задачи

Дано: Частота заболеваний ОРВИ на 1000 человек за год по двум участкам поликлиники. На участке №1 применялись традиционные профилактические лекарственные средства, на участке №2 новые профилактические антивирусные препараты.

месяц	Участок №1	Участок №2
январь	200	130
февраль	280	275
март	302	290
апрель	304	275
май	175	150
июнь	99	60
июль	76	75
август	54	50
сентябрь	60	56
октябрь	78	70
ноябрь	85	87
декабрь	195	156

Требуется:

Определить, эффективно ли новое профилактическое лекарственное средство.

XII вариант

Постановка задачи

Дано: Данные наблюдений, проводимых в двух районах города в течение 20 месяцев, о содержании в воздухе двуокиси углерода в условных единицах (X1-первый район, X2-второй район) приведены в таблице.

Месяц	X1	X2
январь	1	1,3
февраль	1	1,3
март	1,1	1,4
апрель	1,1	1,4
май	1,1	1,5
июнь	1,1	1,5
июль	1	1,1
август	1	1,2
сентябрь	1,2	1,6
октябрь	1,2	1,7
ноябрь	0,6	1
декабрь	0,6	1
январь	0,7	1,1
февраль	0,7	1,15
март	0,75	1,1
апрель	0,7	1,1
май	0,7	1,1
июнь	0,7	1,0
июль	0,8	1,2
август	0,8	1

Требуется:

Определить, справедливо ли утверждение, что содержание в воздухе двуокиси углерода статистически значимо различается в исследуемых районах города.

XIII вариант

Постановка задачи

Дано: Стоимость антисептического средства (руб.) в конкретный момент времени в 8-ми аптеках г. Витебска и г. Могилева.

Номер аптеки	г. Витебск	г. Могилев
1	5,0	4,5
2	5,4	6,0
3	4,7	6,2
4	6,0	4,9
5	3,8	5,7
6	4,8	5,9
7	5,9	5,7
8	4,7	3,4

Требуется:

Определить, справедливо ли утверждение, что стоимость указанного средства в этих городах статистически значимо различаются.

XIV вариант

Постановка задачи

Дано: Количество посещений массажа и водных процедур в поликлинике за определенный период времени

Ежемесячные данные наблюдений
за посещаемостью сеансов массажа и гидротерапии

Месяц	март	апрель	май	июнь	июль	август	сентябрь
Количество посещений массажа	495	503	380	305	348	465	470
Количество посещений водного лечения	132	348	643	865	743	541	480

Требуется:

Определить, есть ли статистически значимые различия в посещении указанных процедур.

Приложение 1

МАТЕРИАЛЫ ДЛЯ ВЫПОЛНЕНИЯ УПРАВЛЯЕМЫХ САМОСТОЯТЕЛЬНЫХ РАБОТ ПОД РУКОВОДСТВОМ ПРЕПОДАВАТЕЛЯ

Примечание.

В заданиях ко всем УСР используются примеры из учебного пособия Медик В.А., Токмачев М.С. «Математическая статистика в медицине»: учеб. пособие / В.А. Медик, М.С. Токмачев – М.: Финансы и статистика,2007 - 800 с. IBSN 978-5-279-03195-5

ПРИМЕНЕНИЕ В ПРОФЕССИОНАЛЬНОЙ ДЕЯТЕЛЬНОСТИ СПЕЦИАЛИСТА ЗДРАВООХРАНЕНИЯ МЕТОДОВ ВЫЯВЛЕНИЯ ЗНАЧИМОСТИ РАЗЛИЧИЙ

ЦЕЛИ ЗАНЯТИЯ

1. Приобрести навыки применения критерия Стьюдента для выявления достоверности различий.
2. Сформировать навыки выявления достоверности различий с помощью непараметрического критерия согласия χ^2 Пирсона.

МЕТОДИКА ВЫПОЛНЕНИЯ РАБОТЫ

Задание 1. Непарный критерий Стьюдента и критерий χ^2 Пирсона

Постановка задачи 1 [5, 302]

Дано: Препарат из группы антагонистов кальция — нифедипин обладает способностью расширять сосуды и его применяют при лечении ишемической болезни сердца. Ш. Хейл и соавторы измеряли диаметр коронарных артерий после приема нифедипина и плацебо и получили две выборки данных диаметра коронарной артерии в мм (таблица 23).

Таблица 23.

Диаметр коронарных артерий при приеме лекарственного средства

плацебо	нифедипин
2,5	2,5
2,2	1,7
2,6	1,5
2	2,5
2,1	1,4
1,8	1,9
2,4	2,3
2,3	2
2,7	2,6
2,7	2,3
1,9	2,2

Требуется:

Определить, позволяют ли приведенные данные считать, что нифедипин значимо влияет на диаметр коронарных артерий.

Решение задачи 1

1. Скопируйте файл с именем «*УСР1 — достоверность различий*» в личную папку.
2. Перейдите на Рабочий лист с именем «*Задание 1*».
3. С помощью соответствующей процедуры пакета Анализ данных **вычислите описательные статистики** двух выборок.
4. Результаты занесите в тетрадь в таблицу 24. В таблице отразите сделанные Вами выводы.

Таблица 24.

Результаты описательной статистики

Статистика	**Выборка**		**Вывод о нормальности в группе (да/нет)**	
	плацебо	**нифедипин**	**плацебо**	**нифедипин**
Среднее				
Медиана				
Мода				
Уровень надежности				
Доверительный интервал			**Есть пересечение ДИ, т.е. наличие общего среднего? (да / нет)**	
Верхняя граница ДИ				
Нижняя граница ДИ				

5. Вычислите с помощью функции ТТЕСТ() (тип 3) вероятность нулевой гипотезы, утверждающей, что средние двух выборок принадлежат одной и той же генеральной совокупности. Результат запишите в таблицу 25.
6. С помощью процедуры пакета Анализ данных ***«Двухвыборочный t-тест с различными дисперсиями»*** вычислите вероятность нулевой гипотезы по критерию Стьюдента для заданных выборок. Сделайте выводы о значимости различий.

Таблица 25.

Результаты выполнения критериев Стьюдента и Пирсона χ^2

Критерий	**PH_0**	**Принимаемая гипотеза (H_0 / H_1)**	**Наличие значимого влияния ЛС на расширение коронарных сосудов сердца (*да/нет*)**	**Правомочно применение метода (*да/нет*)**
Стьюдента (ТТЕСТ())				
Согласия ПИРСОНА (ХИ2ТЕСТ)				

7. Вычислите достоверность различий частотных распределений с помощью критерия согласия Пирсона χ^2 (функция ХИ2ТТЕСТ()), следуя приведенному ниже алгоритму.

Алгоритм применения функции ХИ2ТТЕСТ()

1. Перейдите от признаков к частотам их встречаемости:
   - в каждой выборке данные отсортируйте по возрастанию их значений;
   - выберите и заполните на текущем листе Рабочей книги Microsoft Excel интервалы их встречаемости, учитывая, что значения частот в интервале не должно быть меньше 5;
   - подсчитайте и заполните на Рабочем листе частоты встречаемости признака в каждом интервале.
2. Вычислите ожидаемые значения частот:
   - вычислите суммы частот по каждому признаку и общую сумму частот;
   - вычислите процент частоты по каждому интервалу;
   - вычислите ожидаемые частоты, как произведение процента по каждому интервалу на сумму частот по соответствующей выборке.

3. Вычислите вероятность нулевой гипотезы с помощью функции ХИ2ТЕСТ(), утверждающей что значимые различия между частотными распределениями двух выборок отсутствуют. Вызовите функцию ХИ2ТЕСТ(), В качестве параметров введите интервалы фактических и ожидаемых частот.

4. Результат запишите в таблицу 25. В таблице отразите соответствующие выводы о правомочности применения соответствующего метода и значимости различий.

Задание 2. Парный критерий Стьюдента

Постановка задачи 2 [5, 309]

Дано: Измерена температура у каждого новорожденного из 10 детей в подмышечной впадине (Т1) и в прямой кишке (Т2). Получены следующие значения (таблица 26).

Таблица 26.

Данные о температурах, измеренные на разных участках тела новорожденных

Т1	Т2
36,8	36,9
37,1	37,2
37,3	37,2
37	37,2
37,1	37,3
36,9	37
36,7	36,8
37,2	37,1
37	37,2
36,9	37,1

Требуется:

Определить, можно ли считать, что температура в прямой кишке значимо выше, чем в подмышечной впадине.

Решение задачи 2

1. Перейдите на Рабочий лист с именем «*Задание 2*».
2. С помощью соответствующей процедуры пакета Анализ данных **вычислите описательные статистики** двух выборок. Результаты занесите в тетрадь в таблицу 27. В таблице отразите сделанные Вами выводы.
3. С помощью функции ТТЕСТ() вычислите вероятность нулевой гипотезы о достоверности различий средних по парному критерию Стьюдента (тип=1).

Таблица 27.

Результаты описательной статистики

Статистика	**Выборка**		**Вывод о нормальности в группе (да/нет)**	
	Т1	**Т2**	**Т1**	**Т2**
Среднее				
Медиана				
Мода				
Уровень надежности				
Доверительный интервал			**Есть пересечение ДИ, т.е. наличие общего среднего? (да / нет)**	
Верхняя граница ДИ				
Нижняя граница ДИ				

4. С помощью процедуры пакета Анализ данных «***Парный двухвыборочный t-тест для средних»*** вычислите вероятность нулевой гипотезы двух выборок.
5. Результаты занесите в тетрадь в таблицу 28. В таблице отразите сделанные Вами выводы.

Таблица 28.

Результаты анализа по парному критерию Стьюдента

Критерий	**PH_0**	**Принимаемая гипотеза (H_0 / H_1)**	**Существует ли значимое различие температур (*да/нет*)**	**Правомочно применение метода (*да/нет*)**
Стьюдента (ТТЕСТ())				

Задание 3. Критерий согласия Пирсона χ^2

Постановка задачи 3 [5, 330]

Дано: Результаты исследования органов дыхания пациентов с туберкулезом в течение первого года после заболевания. Рассматриваются две

группы пациентов: группа мужчин — 221 человек, среди них умерших 68; и группа женщин в количестве 194 человека, среди них умерших 83.

Требуется: Установить, значимо ли различается смертность по обеим группам пациентов.

Решение задачи 3

1. Перейдите на Рабочий лист с именем «*Задание 3*».
2. На нем исходные данные преобразованы и представлены в виде таблицы 29.

Таблица 29.

Смертность от туберкулеза органов дыхания среди мужчин и женщин

Группы	Живы	Умерли	Всего
Мужчины	153	68	*
Женщины	111	83	*
Всего	**(*1)**	**(*2)**	**(**)**

3. Вычислите ожидаемые значения для того, чтобы можно было применить функцию ХИ2ТЕСТ(), используя следующий алгоритм.

Алгоритм вычисления ожидаемых значений:

- Вычислите суммы в приведенной таблице по столбцам, строкам и общую сумму, помеченную «**».
- Вычислите значения столбца «Доля в %», как отношение суммы по каждой строке к общей сумме, помеченной «**».
- Вычислите ожидаемые значения, как произведение доли по строке и суммы по соответствующему столбцу **(*1)** или **(*2).**
- Формулы в полученной таблице представлены на рисунке 77.

Расчет ожидаемых значений

					Ожидаемые значения	
Наблюдаемые значения	Живы	Умерли	Всего	Доля в %	Живы	Умерли
Мужчины	153	68	=СУММ(C5:D5)	=E5/$E$7	=G5*$C$7	=G5*$D$7
Женщины	111	83	=СУММ(C6:D6)	=E6/$E$7	=G6*$C$7	=G6*$D$7
Всего	=СУММ(C5:C6)	=СУММ(D5:D6)	=СУММ(C7:D7)	=E7/$E$7	=G7*$C$7	=G7*$D$7

Рис. 77. Формулы для расчета ожидаемых значений

4. Результаты занесите в тетрадь в таблицу 30. В таблице отразите сделанные Вами выводы.

Таблица 30.

Результаты анализа по критерию Пирсона χ2 — функция ХИ2ТЕСТ()

Критерий	PH_0	Принимаемая гипотеза (H_0 / H_1)	Существует ли значимое различие в смертности по группам пациентов (*да/нет*)
Пирсона χ^2 (ХИ2ТЕСТ())			

Задание 4. Критерий согласия Пирсона χ^2

Постановка задачи 4 [5, 330]

Дано: Данные исследования влияния процесса обучения на результаты психологических тестов, проведенные для 100 школьников, в таблице 31.

Таблица 31.

Влияние процесса обучения на психологические тесты

Возраст школьников	Результаты теста		
	Низкие	Средние	Высокие
Младшие	10	15	5
Средние	6	16	8
Старшие	7	13	20

Требуется:

1. Установить наличие влияния обучения на результаты теста.
2. Наличие статистически значимых различий в результатах теста между возрастными группами школьников:
   - Младшие — Средние;
   - Средние — Старшие;
   - Младшие — Старшие.

Решение задачи 4

1. Перейдите на Рабочий лист с именем «*Задание 4*».
2. Используя алгоритм, приведенный в предыдущей задаче, самостоятельно вычислите ожидаемые частоты, результаты разместите в таблице, вид которой приведен на рисунке 78.
3. Вычислите вероятность нулевой гипотезы об отсутствии значимых различий в трех исследуемых группах, применив формулу:

=ХИ2ТЕСТ(B5:D7;H5:J7).

4. Сформулируйте и запишите в тетради нулевую гипотезу, результаты расчета ожидаемых частот (часть таблицы, приведенной на рисунке 78), значение **$Рн_0$(ХИ2ТЕСТ()).**

Влияние процесса обучения на результаты психологического теста

Возраст школьников	**Результаты теста**			**Всего**	**Расчет ожидаемых частот**			
	Низкие	**Средние**	**Высокие**		%	**Низкие**	**Средние**	**Высокие**
Младшие	10	15	5					
Средние	6	16	8					
Старшие	7	13	20					
Всего								

Возрастные группы	**$Рн_0$(ХИ2ТЕСТ())=**	**Наличие значимых отличий (да/нет)**
Для всех возрастных групп		
Младшие - Средние		
Средние - Старшие		
Младшие - Старшие		

Рис. 78. Влияние процесса обучения на результаты психологического теста

5. Запишите вывод о наличии значимого влияния обучения на результат теста, применив поправку Бонферрони (α/k, где α — уровень значимости; k — количество сравнений групп).

6. Вычислите значения **$Рн_0$(ХИ2ТЕСТ())** для возрастных групп, задавая соответствующие значения фактического и ожидаемого интервалов:

- Младшие — Средние;
- Средние — Старшие;
- Младшие — Старшие.

7. В тетради запишите результаты расчета ожидаемых частот (правая часть таблицы, приведенной на рисунке 78), значения **$Рн_0$(ХИ2ТЕСТ())=** при сравнении **разных возрастных групп** и соответствующие им выводы **о наличии** или **отсутствии различий.**

✓ **Обратите внимание!** Перед использованием функции ХИ2ТЕСТ() при необходимости можно копировать нужные интервалы данных в свободные ячейки таблицы Рабочего листа Excel.

8. Результаты работы сохраните в личной папке.

ПРИМЕНЕНИЕ КОРРЕЛЯЦИОННОГО И РЕГРЕССИОННОГО АНАЛИЗОВ ПРИ РЕШЕНИИ ЗАДАЧ МЕДИЦИНЫ И ЗДРАВООХРАНЕНИЯ

ЦЕЛИ ЗАНЯТИЯ

1. Приобрести навыки практического применения программных средств, позволяющих выявить взаимосвязи между признаками.
2. Сформировать навыки применения корреляционного анализа в профессиональной деятельности.
3. Сформировать навыки практического проведения регрессионного анализа.

МЕТОДИКА ВЫПОЛНЕНИЯ РАБОТЫ

Корреляционный анализ

Постановка задачи 1 [5, 302]

Дано: Рост отцов и их взрослых сыновей в таблице 32.

Требуется: Выявить наличие связи между ростом отцов и ростом взрослых сыновей.

Таблица 32.

Данные роста отцов и их взрослых сыновей

Рост отцов	Рост сыновей
180	186
172	180
173	176
169	171
175	182
170	166
179	182
170	172
167	169
174	177

Постановка задачи 2 [5, 425]

Дано: Основные показатели по фтизиатрической службе по 14 районам; смертность, заболеваемость, болезненность и эффективность медосмотров в таблице 33. При этом: ***смертность*** — число больных, умерших в течение года от активных форм туберкулеза и состоявших на учете по данной территории на 100 тыс. среднегодовой численности населения; ***заболеваемость*** — число больных со всеми формами активного туберкулеза, выявленных впервые в отчетном году на 1000 тыс. среднегодовой численности населения; ***болезненность*** — число больных всеми формами активного туберкулеза, состоящих на учете на конец отчетного года на 100 тыс. населения, ***эффективность медосмотров*** — процент впервые выявленных боль-

ных. Показатель смертности принят в качестве результирующего признака (Y), а заболеваемость (X_1), болезненность (X_2), эффективность медосмотров (X_3), рассматриваются в качестве факторов, участвующих в формировании показателей смертности.

Требуется: Выявить наличие связи и ее силу между показателем смертности (Y) и заболеваемостью (X_1), болезненностью (X_2), эффективностью медосмотров (X_3), которые рассматриваются в качестве факторов, участвующих в формировании показателей смертности.

Регрессионный анализ

Постановка задачи 3 [5, 425]

Дано: Рост отцов и их взрослых сыновей в таблице 32.

Требуется: Выявить наличие функциональной линейной связи между ростом отцов и ростом взрослых сыновей.

Постановка задачи 4 [5, 425]

Дано: Основные показатели по фтизиатрической службе по 14 районам; смертность, заболеваемость, болезненность и эффективность медосмотров в таблице 33.

Требуется: Выявить функциональную зависимость между показателем смертности (Y) и заболеваемостью (X_1), болезненностью (X_2), эффективностью медосмотров (X_3), которые рассматриваются в качестве факторов, участвующих в формировании показателей смертности.

Таблица 33.

Значение основных показателей по фтизиатрической службе

Район	Смертность на 100 тыс. населения чел.	Заболеваемость на 1000 тыс. населения чел.	Болезненность на 100 тыс. населения чел.	Эффективность медосмотров
	Y	X_1	X_2	X_3
1	13,3	76,4	281,9	47,7
2	11,8	53,3	278,3	71,4
3	12,4	62,2	273,1	50,1
4	9,2	73,3	214,2	57,3
5	17,2	72,7	294,5	40,3
6	10,7	53,7	254,1	61,3
7	12,3	55,6	253,3	47,8
8	24,4	95,4	301,8	39,2
9	16,3	73,7	285,3	39,4
10	10	55,3	236,2	63,6
11	11,4	75,8	231,8	72,7
12	23,1	89,1	279,4	28,5
13	12	52,6	275,4	58,6
14	9,8	50,6	248,8	58,6

Корреляционный анализ

ХОД ВЫПОЛНЕНИЯ ЗАДАНИЯ 1

I. Описательные статистики

1. Перейдите на Рабочий лист с именем «*Задание 1-1*».
2. С помощью соответствующей процедуры пакета Анализ данных **вычислите описательные статистики** выборок по столбцам: рост отцов и рост сыновей, выполните проверку частотных распределений выборок на нормальность.
3. Результаты занесите в тетрадь в таблицу 34. В таблице отразите сделанные Вами выводы. Определите, какие методы следует применять в данном случае для вычисления коэффициента корреляции (параметрические или непараметрические).

Таблица 34.

Основные показатели описательной статистики и выводы

Статистика	**Выборка**	
	Рост отцов	**Рост сыновей**
Среднее		
Медиана		
Мода		
Уровень надежности		
Вывод о нормальности частотного распределения (да/нет)		
Метод анализа (параметрический, непараметрический)		

II. Корреляция Пирсона (параметрический метод)

1. Используя параметрические методы, вычислите с помощью функции ***КОРРЕЛ()*** коэффициент линейной корреляции. Сделайте вывод о направлении и силе связи между ростом отцов и их взрослых сыновей. Результат запишите в таблицу 35.
2. Перейдите на Рабочий лист *«Задание 1-2»*.
3. Вычислите значение коэффициента корреляции Пирсона с помощью процедуры «***Корреляция»*** пакета Анализ данных. Результат запишите в таблицу 35.

Таблица 35.

Коэффициенты корреляции, рассчитанные различными методами

Способ определения R	**R**	**Сила связи между признаками: *сильная, средняя, умеренная, слабая***	**Направление связи: *прямая, обратная***	**Правомочно применение метода (*да/нет*)**
Функция **КОРРЕЛ()**				
Процедура **«КОРРЕЛЯЦИЯ»**				
Корреляция **СПИРМЕНА**				

III. Корреляция Спирмена (непараметрический метод)

1. Перейдите на Рабочий лист *«Задание 1-3»*, вычислите непараметрический коэффициент корреляции ρ, используя формулу Спирмена:

$$\rho=1-6*\sum d_i^2/ (n*(n^2-1)) \qquad (1)$$

2. Для вычисления коэффициента корреляции ρ примените следующий алгоритм.

Алгоритм вычисления коэффициента корреляции Спирмена:

- вычислите ранги по каждой выборке, используя процедуру «***Ранг и персентиль»*** из пакета Анализ данных;
- отсортируйте по **каждому признаку** данные о рангах по возрастанию поля «Точка»;
- вычислите для каждой точки разность рангов $\mathbf{d_i}$;
- вычислите для каждой точки квадрат разности рангов $\mathbf{d_i^2}$;
- вычислите сумму квадратов разности рангов;
- вычислите значение **ρ** по формуле (1);
- результат запишите в таблицу 35;
- сделайте вывод о направлении и силе связи;
- отметьте в таблице 35, какой метод является в данном случае адекватным.

ХОД ВЫПОЛНЕНИЯ ЗАДАНИЯ 2

I. Описательные статистики

1. Перейдите на Рабочий лист с именем «*Задание 2-1*».
2. С помощью соответствующей процедуры пакета Анализ данных **вычислите описательные статистики** всех выборок Y, X_1, X_2, X_3, выполните проверку на нормальность частотных распределений выборок.
3. Результаты занесите в тетрадь в таблицу 36. В таблице отразите сделанные Вами выводы. Определите, какие методы следует применять в данном случае для вычисления коэффициента корреляции (параметрические или непараметрические).

Таблица 36.

Основные показатели описательной статистики и выводы

Статистика	**Выборка**			
	Y	**X_1**	**X_2**	**X_3**
Среднее				
Медиана				
Мода				
Уровень надежности				
Вывод о нормальности частотного распределения (да/нет)				
Метод анализа (параметрический, непараметрический)				

II. Корреляция Пирсона (параметрический метод)

1. Вычислите значение коэффициента корреляции Пирсона с помощью процедуры «***Корреляция»*** пакета Анализ данных. Результат запишите в таблицу 37.

Таблица 37.

Коэффициенты корреляции, рассчитанные различными методами

Метод определения R	**R Y-X_1**	**Сила/ направ-ление связи**	**R Y-X_2**	**Сила/ направ-ление связи**	**R Y-X_3**	**Сила/ направ-ление связи**	**Право-мочен метод (*да/нет)***
Процедура **«КОРРЕЛЯЦИЯ»**							
Корреляция **СПИРМЕНА**							

III. Корреляция Спирмена (непараметрический метод)

1. Перейдите на Рабочий лист с именем *«Задание 2-2»*, вычислите непараметрические коэффициенты корреляции ρ, определяющие силу связи между признаками ***Y-X_1, Y-X_2, Y-X_3***. Для вычисления ρ используйте выше приведенный алгоритм и формулу Спирмена (1).
2. Запишите полученные результаты в таблицу 37.
3. Сравните результаты, полученные разными методами.

Регрессионный анализ

ХОД ВЫПОЛНЕНИЯ ЗАДАНИЯ

Решение задачи 3

1. Перейдите на Рабочий лист с именем «*Задание 3*».
2. С помощью процедуры «***Регрессия»*** пакета Анализ данных **проведите регрессионный анализ** данных**, позволяющий определить** уравнение связывающее рост сыновей (Y) с ростом отцов (X), считая значение **a_0 не равным 0**. Результаты и сделанные Вами выводы занесите в тетрадь в таблицу 38.
3. Предусмотрите при выполнении процедуры вывод графика зависимости Y=f(X). Обратите внимание, является ли он линейным.

Таблица 38.

Результаты регрессионного анализа

Решение при a_0	**a_0**	**p a_0**	**a_1**	**p a_1**	**R**	**R-квад-рат**	**Линей-ность (да/нет)**	**Знач. F**	**Значима модель (да/нет)**	**Уравнение**
$a_0 \neq 0$										
$a_0=0$	-	-								

4. Используя процедуру «***Регрессия»*** пакета Анализ данных **проведите регрессионный анализ** данных**, позволяющий определить** уравнение связывающее рост сыновей (Y) с ростом отцов (X), считая значение **a_0 равным 0**. Результаты и сделанные Вами выводы занесите в тетрадь в таблицу 38. Запишите уравнение, характеризующее зависимость роста сыновей (Y) от роста отцов (X).

5. Предусмотрите при выполнении процедуры вывод графика зависимости Y=f(X). Обратите внимание, является ли он линейным.

6. Составьте прогноз, каким будет рост взрослого сына, если рост отца равен 188 см.

Решение задачи 4

1. Перейдите на Рабочий лист с именем «*Задание 4*».

2. С помощью процедуры «***Регрессия»*** пакета Анализ данных **проведите регрессионный анализ** данных**, позволяющий определить** уравнение**,** описывающее зависимость признака смертность (Y) от признаков заболеваемость (X_1), болезненность (X_2), эффективность медосмотров (X_3), считая значение **a_0 не равным 0.** Результаты занесите в тетрадь в таблицу 39. В таблице отразите сделанные Вами выводы.

3. Предусмотрите при выполнении процедуры вывод графика зависимости Y=f(X). Обратите внимание, является ли он линейным.

Таблица 39.

Результаты регрессионного анализа

a_0	a_0	P a_0	a_1	P a_1	a_2	P a_2	a_3	P a_3	R	R-квадрат	Линейность (да/нет)	Знач. F	Значима модель (да/нет)	Уравнение
$a_0 \neq 0$														
$a_0=0$	-	-												

4. Используя процедуру «***Регрессия»*** пакета Анализ данных **проведите регрессионный анализ** данных**, позволяющий определить** уравнение**,** описывающее зависимость признака смертность (Y) от признаков заболеваемость (X_1), болезненность (X_2), эффективность медосмотров (X_3), считая значение **a_0 равным 0**. Результаты занесите в тетрадь в таблицу 39. В таблице отразите сделанные Вами выводы. Запишите уравнение, отражающее линейную зависимость значения признака смертность (Y) от признаков заболеваемость (X_1), болезненность (X_2), эффективность медосмотров (X_3).

5. Предусмотрите при выполнении процедуры вывод графика зависимости Y=f(X). Обратите внимание, является ли он линейным.

6. Используя полученное уравнение, проанализируйте какие из факторов (X_1, X_2, X_3) повышают уровень смертности Y, а какие понижают? Определите степень влияния каждого фактора на уровень смертности.

ПРИМЕНЕНИЕ ОДНОФАКТОРНОГО И ДВУХФАКТОРНОГО ДИСПЕРСИОННОГО АНАЛИЗОВ ДЛЯ ВЫЯВЛЕНИЯ ВЛИЯНИЯ РЯДА ФАКТОРОВ НА ХОД ПРОФЕССИОНАЛЬНО ЗНАЧИМЫХ ПРОЦЕССОВ

ЦЕЛИ ЗАНЯТИЯ

1. Приобрести навыки практического применения процедур пакета Анализа Microsoft Excel, позволяющих выявить значимое влияние факторов на результирующий признак.

2. Сформировать навыки применения однофакторного дисперсионного анализа в профессиональной деятельности.

3. Сформировать навыки практического применения двухфакторного дисперсионного анализа для решения задач медицины и здравоохранения.

МЕТОДИКА ВЫПОЛНЕНИЯ РАБОТЫ

Однофакторный дисперсионный анализ

Постановка задачи 1 [5, 359]

Дано: Из данных о количестве заболеваний органов дыхания за два года среди взрослого населения определенной возрастной категории случайным образом были отобраны три группы по 4 человека каждая. Из них: 1 группа не курящие; 2 группа — стаж курильщика до 5 лет, 3 группа — стаж курильщика более 5 лет, данные представлены в таблице 40.

Требуется:

1) Выявить наличие влияния длительности курения на заболевание органов дыхания;

2) Вычислить выборочный коэффициент детерминации, который показывает, какая доля выборочной дисперсии результативного признака (количество заболеваний органов дыхания) объясняется его зависимостью от влияющего фактора (длительность курения).

Таблица 40.

Количество заболеваний органов дыханий у курящих

Не курящие	Курящие меньше 5 лет	Курящие 5 лет и более
1	3	3
0	2	4
1	2	5
2	1	3

Постановка задачи 2 [5, 364]

Дано: Масса и рост тела 20 студентов. В качестве отклика (результирующего признака) полагаем массу тела. Рост тела, фактор, влияющий на

массу тела, представим в виде 5-ти уровней: 155-160 см; 160-165 см; 165-170 см; 170-175 см; 175-180 см. Данные представлены в таблице 41.

Требуется:

1) Выявить наличие значимости влияния роста студента на его массу.
2) Вычислить выборочный коэффициент детерминации.

Таблица 41.

Данные зависимости массы студентов от их роста

Номер испытания	Уровни фактора (см)				
	155-160	160-165	165-170	170-175	175-180
1	59	60	63	67	73
2	53	54	61	68	79
3		57	68	74	71
4		59	62	72	75
5			64	69	

Двухфакторный дисперсионный анализ с повторениями

Задание 3

Постановка задачи 3 [5, 382]

Дано: Результаты исследования работоспособности шести групп детей с помощью психологического теста приведены в таблице 42. Одним из показателей, с помощью которого определяется работоспособность, является количество ошибок, допускаемых каждым из детей за единицу времени. Экспериментальным путем исследовали влияние двух факторов: А — изменения работоспособности во время учебных занятий и В — изменения работоспособности в зависимости от возраста. Фактор А имеет три уровня: «до учебных занятий», «во время учебных занятий», «после учебных занятий». Фактор В имеет два уровня: «дети старшего возраста», «дети младшего возраста».

Таблица 42.

Данные психологического теста работоспособности

Факторы	старшие дети	младшие дети
до занятий	2	2
	0	4
	3	4
	4	6
во время занятий	2	0
	3	5
	5	6
	6	7
после занятий	4	7
	5	8
	6	9
	7	9

Требуется определить:

1) снижается ли работоспособность во время учебного процесса вследствие наступившего утомления;

2) находится ли работоспособность школьников в зависимости от возраста;

3) оказывает ли влияние комбинированное воздействие на работоспособность факторов утомляемости и возраста.

4) выборочные коэффициенты детерминации для каждого фактора и их взаимного влияния.

Двухфакторный дисперсионный анализ без повторений

Постановка задачи 4 [5, 383]

Дано: Результаты исследования причин смертности в разных социальных группах населения Франции за 1986 г. (на 100 тыс. населения) приведены в таблице 43.

Требуется определить влияние:

1) профессионального фактора на смертность;

2) заболеваемости и бытового фактора на смертность;

3) комбинированного воздействия факторов профессиональной принадлежности и бытового фактора на смертность;

4) выборочного коэффициента детерминации для каждого фактора.

Таблица 43.

Данные о причинах смертности различных социальных групп населения

Причина смертности	Вид занятий				
	Руководители высшего звена	Преподаватели	Руководители среднего звена	Сельскохозяйственные рабочие	Промышленные рабочие
Новообразования	150	140	205	290	350
Сердечно-сосудистые болезни	130	150	180	190	185
Насчастные случаи	45	30	75	175	95
Цирроз печени	15	16	33	75	95
Самоубийства	20	25	36	30	45

ХОД ВЫПОЛНЕНИЯ ЗАДАНИЙ

I. Однофакторный дисперсионный анализ

Решение задачи 1

В связи с тем, что количество объектов во всех группах исходных данных (таблица 40) одинаковое, можно не проверять нормальность частотных распределений и равенство дисперсий в выборках, а сразу применить дисперсионный анализ.

1. Перейдите на Рабочий лист с именем «*Задание 1*».
2. Вызовите из пакета Анализ данных процедуру «***Однофакторный дисперсионный анализ».***
3. Задайте:
   - в качестве входного интервала диапазон ячеек ***B$6:$D$10***;
   - установку флажка «***Метки***»;
   - ***альфа*** равное ***0,05***;
   - начало выходного интервала с ячейки ***$C$24.***
4. Запишите основные результаты дисперсионного анализа в таблице 44.

Таблица 44.

Результаты однофакторного дисперсионного анализа

Источник вариации	**SS**	**P-значение**	**Влияние фактора значимо (да/нет)**	**ρ**
Между группами				
Внутри групп				
Итого				

5. Сделайте вывод, является ли значимыми результаты данной модели исследования.
6. Вычислите выборочный коэффициент детерминации:

ρ= SS между группами/ SS общая (в строке итого).

7. Сделайте вывод о том, какой процент изменчивости признака (количество заболеваний органов дыхания) объясняется влиянием фактора (стаж курения). Вывод запишите в тетрадь.

Решение задачи 2

1. Перейдите на Рабочий лист *«Задание 2».*
2. Обоснуйте правомерность использования дисперсионного анализа, учитывая, что количество данных в выборках различно. Для этого проверьте на нормальность частотные распределения выборок, предварительно вычислив описательные статистики каждой выборки. В качестве критерия нормальности используйте равенство среднего и медианы.
3. Самостоятельно примените однофакторный дисперсионный анализ для определения влияния роста студента на массу его тела.

4. Запишите основные результаты дисперсионного анализа в табл. 45.

Таблица 45

Результаты однофакторного дисперсионного анализа

Источник вариации	**SS**	**P-значение**	**Влияние фактора значимо (да/нет)**	**ρ**
Между группами				
Внутри групп				
Итого				

5. Сделайте вывод, являются ли значимыми результаты данной модели исследования.

6. Вычислите выборочный коэффициент детерминации ρ.

7. Сделайте вывод о том, какой процент изменчивости признака (масса тела) объясняется влиянием фактора (рост студента). Вывод запишите в тетрадь.

II. Двухфакторный дисперсионный анализ с повторениями

Решение задачи 3

1. Перейдите на Рабочий лист с именем «*Задание 3*».
2. Так как во всех элементарных выборках одинаковое количество данных, проверять нормальность их частотных распределений не надо.
3. Вызовите процедуру «***Двухфакторный дисперсионный анализ с повторениями»*** из пакета Анализ данных.
4. Задайте следующие параметры в одноименном окне процедуры:
   - входной интервал: **$B$6:$D$18,**
   - число строк для выборки: **4,**
   - альфа: **0,05,**
   - выходной интервал: **F6**.
5. Результаты запишите в таблицу 46.

Таблица 46

Результаты двухфакторного дисперсионного анализа с повторениями

Источник вариации	**SS**	**P-значение**	**Значимо влияние фактора (да/нет)**	**ρ**
Выборка ***(когда?)***				
Столбцы ***(кто?)***				
Взаимодействие ***(когда*кто)***				
Итого		-	-	-

6. Сделайте выводы о значимости влияния факторов и их взаимодействия, запишите выводы в таблицу 46.

7. Вычислите выборочные коэффициенты детерминации для каждого фактора и их взаимодействия, запишите их значения в таблицу 46.

8. Сделайте выводы о том, какая доля изменчивости результирующего признака объясняется изменчивостью каждого фактора.

III. Двухфакторный дисперсионный анализ без повторений

Решение задачи 4

1. Перейдите на Рабочий лист с именем «*Задание 4*».

2. Так как во всех элементарных выборках одинаковое количество данных, проверять нормальность их частотных распределений не надо.

3. Вызовите процедуру «***Двухфакторный дисперсионный анализ без повторений»*** из пакета Анализ данных.

4. Задайте следующие параметры в одноименном окне процедуры:

- входной интервал: **$B$6:$G$11;**
- **установите флажок «Метки»;**
- альфа: **0,05;**
- выходной интервал: **К5**.

5. Результаты запишите в таблицу 47.

Таблица 47

Результаты двухфакторного дисперсионного анализа без повторений

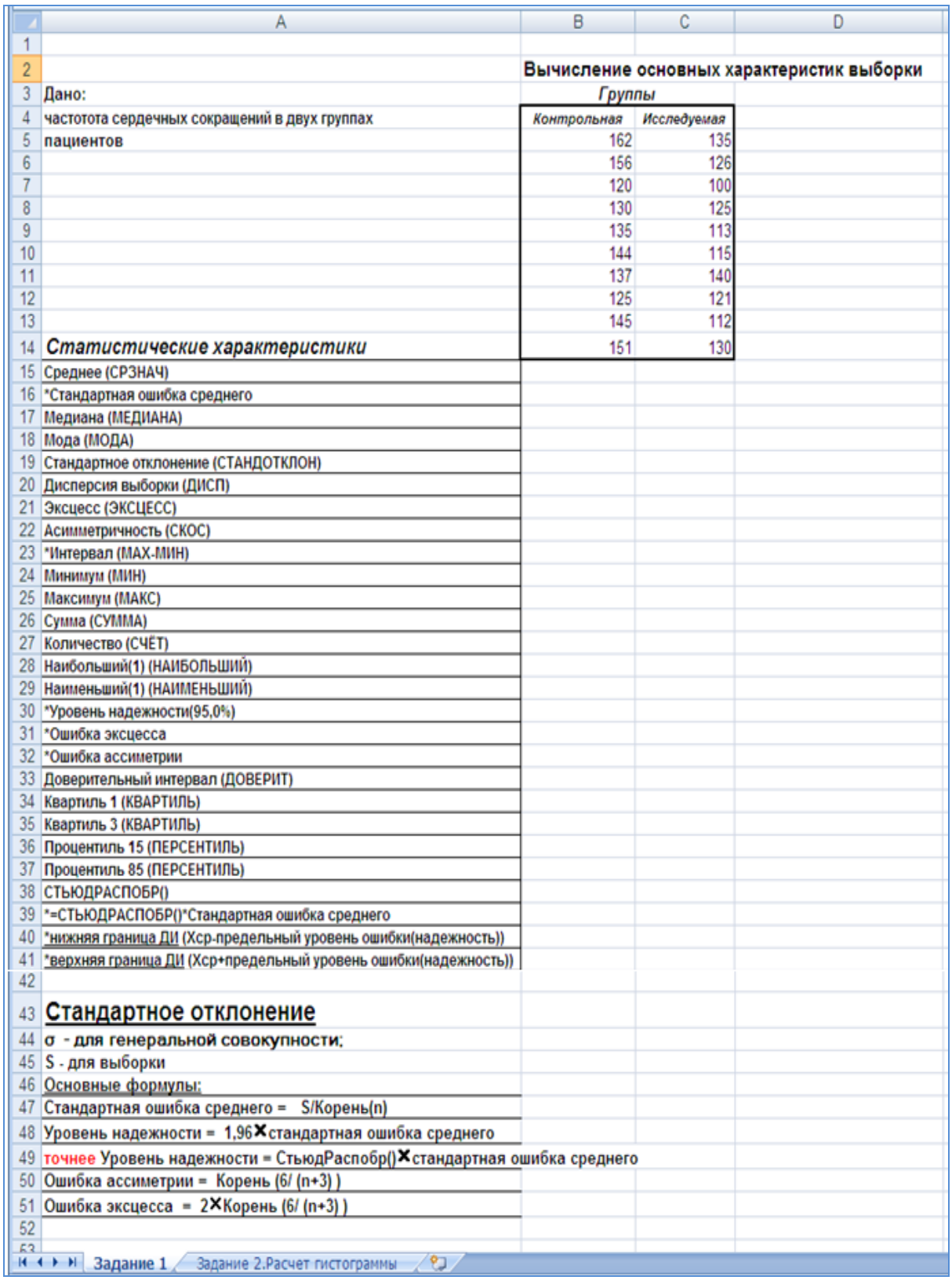
Источник вариации	**SS**	**P-значение**	**Значимо влияние фактора (да/нет)**	**ρ**
Строки (причина смертности)				
Столбцы (вид занятий)				
Итого		-	-	-

6. Вычислите выборочные коэффициенты детерминации для каждого фактора, запишите их значения в таблицу 46.

7. Сделайте выводы о том, какая доля изменчивости результирующего признака объясняется изменчивостью каждого фактора.

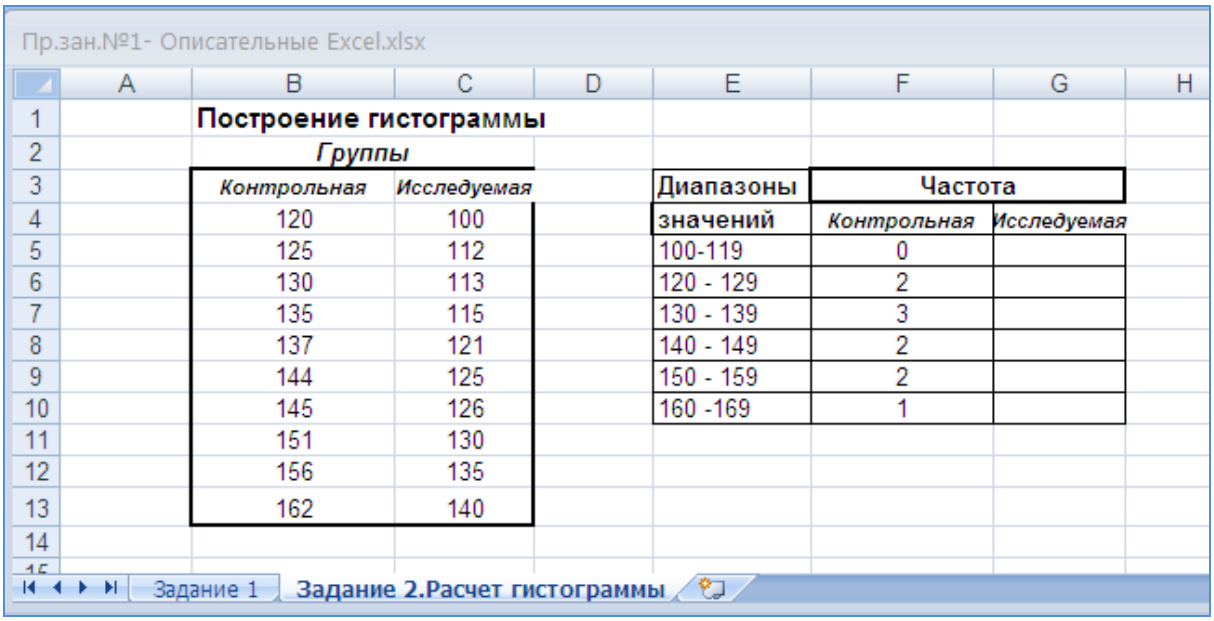
МАТЕРИАЛЫ ЗАДАНИЙ ДЛЯ ВЫПОЛНЕНИЯ ПРАКТИЧЕСКИХ РАБОТ

ПРИМЕНЕНИЕ СТАТИСТИЧЕСКИХ ФУНКЦИЙ MICROSOFT EXCEL ДЛЯ ВЫЧИСЛЕНИЯ ОСНОВНЫХ СТАТИСТИЧЕСКИХ ХАРАКТЕРИСТИК ВЫБОРКИ. ЧАСТОТНЫЙ АНАЛИЗ ДАННЫХ



	A	B	C	D
1				
2		**Вычисление основных характеристик выборки**		
3	Дано:	*Группы*		
4	частотота сердечных сокращений в двух группах	*Контрольная*	*Исследуемая*	
5	пациентов	162	135	
6		156	126	
7		120	100	
8		130	125	
9		135	113	
10		144	115	
11		137	140	
12		125	121	
13		145	112	
14	***Статистические характеристики***	151	130	
15	Среднее (СРЗНАЧ)			
16	*Стандартная ошибка среднего			
17	Медиана (МЕДИАНА)			
18	Мода (МОДА)			
19	Стандартное отклонение (СТАНДОТКЛОН)			
20	Дисперсия выборки (ДИСП)			
21	Эксцесс (ЭКСЦЕСС)			
22	Асимметричность (СКОС)			
23	*Интервал (МАХ-МИН)			
24	Минимум (МИН)			
25	Максимум (МАКС)			
26	Сумма (СУММА)			
27	Количество (СЧЁТ)			
28	Наибольший(1) (НАИБОЛЬШИЙ)			
29	Наименьший(1) (НАИМЕНЬШИЙ)			
30	*Уровень надежности(95,0%)			
31	*Ошибка эксцесса			
32	*Ошибка ассиметрии			
33	Доверительный интервал (ДОВЕРИТ)			
34	Квартиль 1 (КВАРТИЛЬ)			
35	Квартиль 3 (КВАРТИЛЬ)			
36	Процентиль 15 (ПЕРСЕНТИЛЬ)			
37	Процентиль 85 (ПЕРСЕНТИЛЬ)			
38	СТЬЮДРАСПОБР()			
39	*=СТЬЮДРАСПОБР()*Стандартная ошибка среднего			
40	*нижняя граница ДИ (Хср-предельный уровень ошибки(надежность))			
41	*верхняя граница ДИ (Хср+предельный уровень ошибки(надежность))			
42				
43	**Стандартное отклонение**			
44	**σ - для генеральной совокупности;**			
45	S - для выборки			
46	Основные формулы:			
47	Стандартная ошибка среднего = S/Корень(n)			
48	Уровень надежности = 1,96✕стандартная ошибка среднего			
49	точнее Уровень надежности = СтьюдРаспобр()✕стандартная ошибка среднего			
50	Ошибка ассиметрии = Корень (6/ (n+3))			
51	Ошибка эксцесса = 2✕Корень (6/ (n+3))			
52				

Задание 1 | Задание 2.Расчет гистограммы

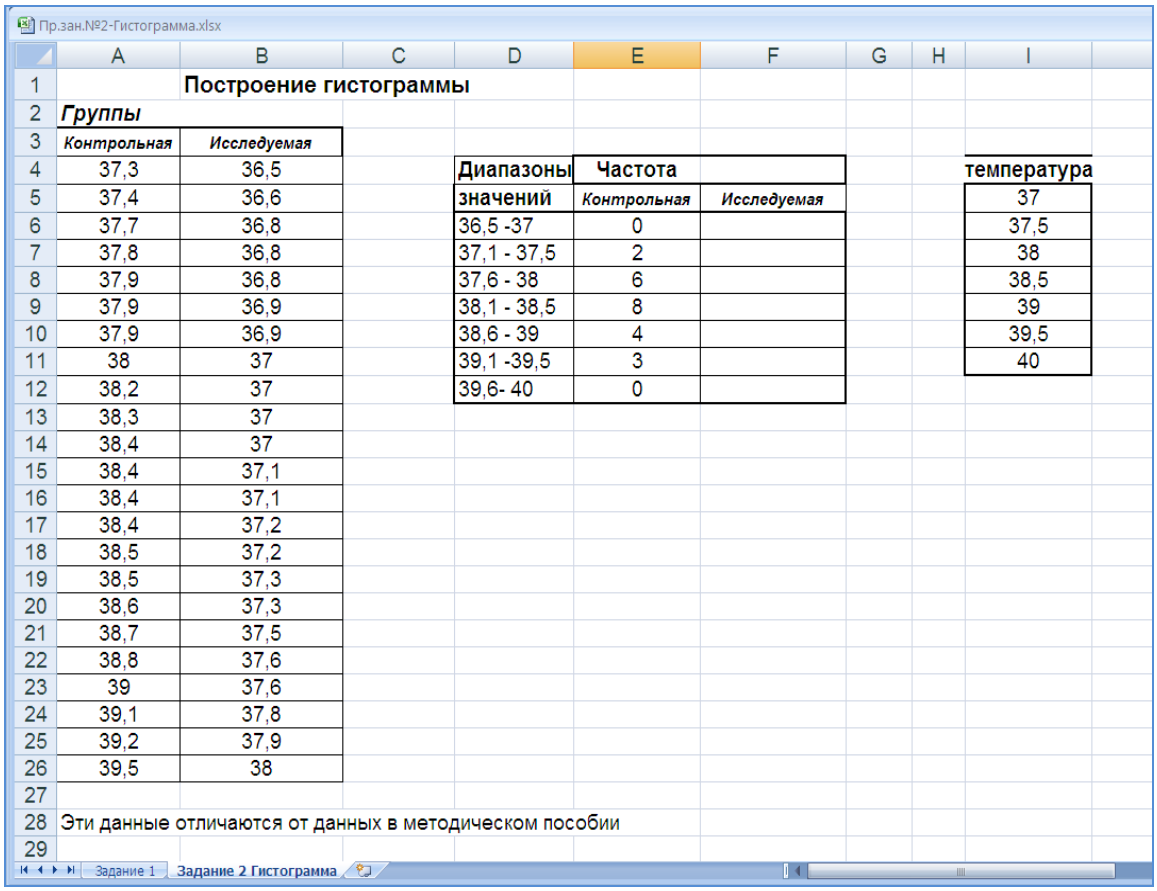

Пр.зан.№1- Описательные Excel.xlsx
Построение гистограммы
Группы
Контрольная
Исследуемая
120
125
130
135
137
144
145
151
156
162
100
112
113
115
121
125
126
130
135
140
Диапазоны значений
Частота
Контрольная
Исследуемая
100-119
120 - 129
130 - 139
140 - 149
150 - 159
160 -169
0
2
3
2
2
1
Задание 1
Задание 2.Расчет гистограммы

ВОЗМОЖНОСТИ ПАКЕТА АНАЛИЗ ДАННЫХ.
ОПИСАТЕЛЬНАЯ СТАТИСТИКА,
ГИСТОГРАММА ЧАСТОТНОГО РАСПРЕДЕЛЕНИЯ ВЫБОРКИ

Пр.зан.№2-Гистограмма.xlsx

	A	B	C	D
1				
2	**Вычисление основных характеристик выборки**			
3	**Дано:**	***Группы***		
4	**температура пациентовй в двух группах**	***Контрольная***	***Исследуемая***	
5		38,5	37,6	
6		38,2	37,1	
7		39	38,1	
8		39,5	38,2	
9		38,7	38	
10		37	37,9	
11		38,4	36,8	
12		38,3	37,1	
13		39,2	38,2	
14		37,9	36,8	
15		37,9	36,5	
16		38,4	38,1	
17		38,6	38,2	
18		38,4	36,7	
19		37,3	36,9	
20		37,7	36,8	
21		37,4	37	
22		37	37,5	
23		38,5	37,6	
24		37,9	37,2	
25		37,8	37,3	
26		38	37,6	
27		38,8	38,2	
28	***Статистические характеристики***			
29	**Среднее (СРЗНАЧ)**			
30	***Стандартная ошибка среднего**			
31	**Медиана (МЕДИАНА)**			
32	**Мода (МОДА)**			
33	**Стандартное отклонение(СТАНДОТКЛОН)**			
34	**Дисперсия выборки (ДИСП)**			
35	**Эксцесс (ЭКСЦЕСС)**			
36	**Асимметричность (СКОС)**			
37	***Интервал (МАХ-МИН)**			
38	**Минимум (МИН)**			
39	**Максимум (МАКС)**			
40	**Сумма (СУММ)**			
41	**Количество (СЧЕТ)**			
42	**Наибольший(1) (НАИБОЛЬШИЙ)**			
43	**Наименьший(1) (НАИМЕНЬШИЙ)**			
44	***Уровень надежности(95,0%)**			
45	***Ошибка эксцесса**			
46	***Ошибка ассиметрии**			
47	**Доверительный интервал (ДОВЕРИТ)**			
48	**Квартиль 1 (КВАРТИЛЬ)**			
49	**Квартиль 3 (КВАРТИЛЬ)**			
50	**Процентиль 15 (ПЕРСЕНТИЛЬ)**			
51	**Процентиль 85 (ПЕРСЕНТИЛЬ)**			
52	**СТЬЮДРАСПОБР()**			
53	***СТЬЮДРАСПОБР()*Стандартная ошибка среднего**			
54	***нижняя граница ДИ** (Хср-предельный уровень ошибки(надежность))			
55	***верхняя граница ДИ** (Хср+предельный уровень ошибки(надежность))			
56	**Стандартное отклонение**			
57	**σ - для генеральной совокупности;**			
58	**S - для выборки**			
59	**Основные формулы:**			
60	**Стандартная ошибка среднего = S/Корень(n)**			
61	**Уровень надежности = 1,96✕стандартная ошибка среднего**			
62	**точнее Уровень надежности = СтьюдРаспобр()✕стандартная ошибка среднего**			
63	**Ошибка ассиметрии = Корень (6/ (n+3))**			
64	**Ошибка эксцесса = 2✕Корень (6/ (n+3))**			
65				

Задание 1 | Задание 2 Гистограмма



Пр.зан.№2-Гистограмма.xlsx

	A	B	C	D	E	F	G	H	I
1		**Построение гистограммы**							
2	***Группы***								
3	***Контрольная***	***Исследуемая***							
4	37,3	36,5		**Диапазоны**	**Частота**				**температура**
5	37,4	36,6		**значений**	***Контрольная***	***Исследуемая***			37
6	37,7	36,8		36,5 -37	0				37,5
7	37,8	36,8		37,1 - 37,5	2				38
8	37,9	36,8		37,6 - 38	6				38,5
9	37,9	36,9		38,1 - 38,5	8				39
10	37,9	36,9		38,6 - 39	4				39,5
11	38	37		39,1 -39,5	3				40
12	38,2	37		39,6- 40	0				
13	38,3	37							
14	38,4	37							
15	38,4	37,1							
16	38,4	37,1							
17	38,4	37,2							
18	38,5	37,2							
19	38,5	37,3							
20	38,6	37,3							
21	38,7	37,5							
22	38,8	37,6							
23	39	37,6							
24	39,1	37,8							
25	39,2	37,9							
26	39,5	38							
27									
28	Эти данные отличаются от данных в методическом пособии								
29									

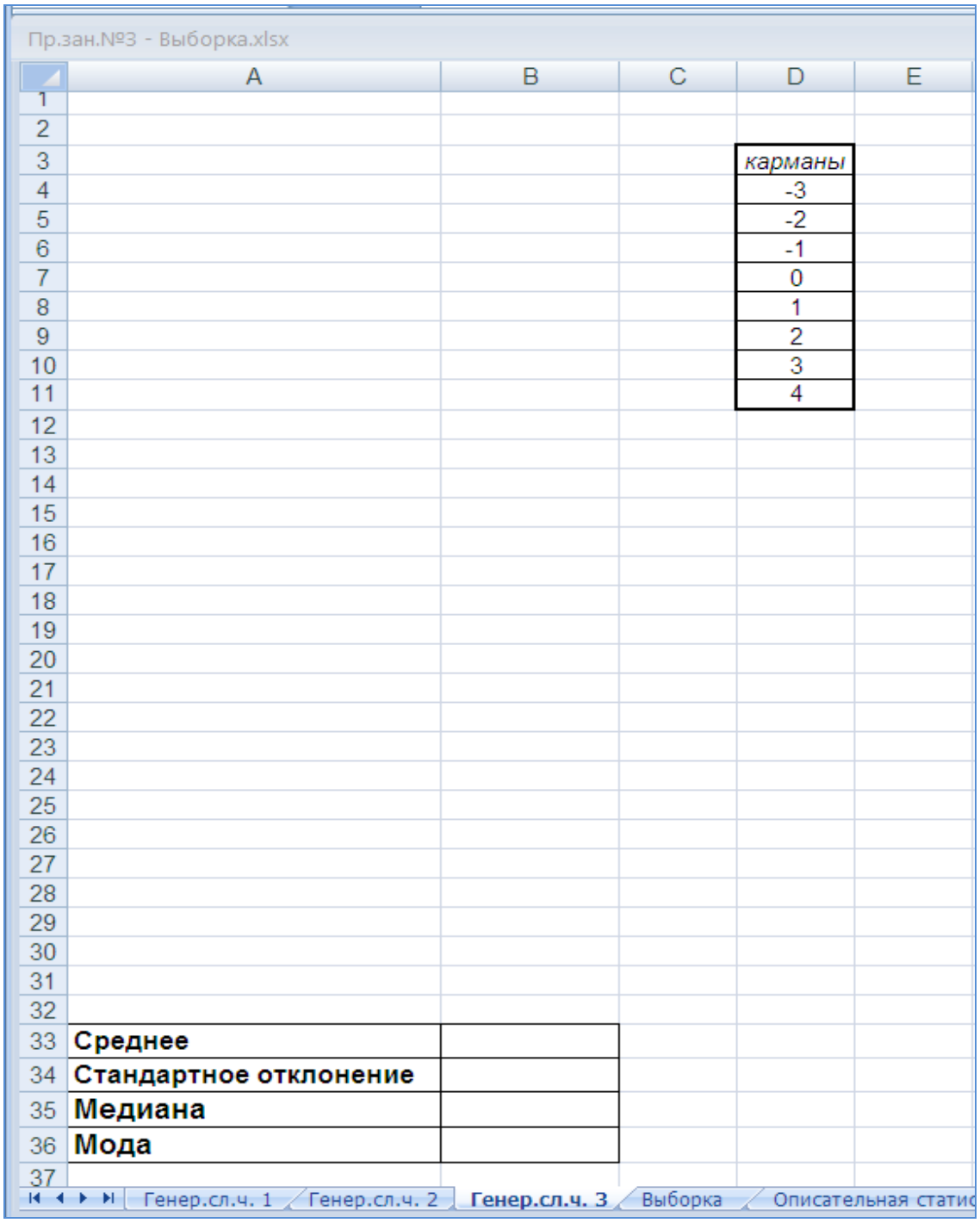
Задание 1 | Задание 2 Гистограмма

ФОРМИРОВАНИЕ РАНДОМИЗИРОВАННОЙ ВЫБОРКИ И ИЗУЧЕНИЕ ЕЕ СВОЙСТВ С ПОМОЩЬЮ ИНСТРУМЕНТОВ ПАКЕТА АНАЛИЗ ДАННЫХ

Пр.зан.№3 - Выборка.xlsx

	A	B	C	D	E	F	G
1		План эксперимента					
2	Номер образца	Вероятнсть использования образца	проверка, сколько раз в месяц использовался образец		номер дня месяца	номер, используемого образца	
3	1	0,1			1		
4	2	0,1			2		
5	3	0,1			3		
6	4	0,1			4		
7	5	0,1			5		
8	6	0,1			6		
9	7	0,1			7		
10	8	0,1			8		
11	9	0,1			9		
12	10	0,1			10		
13		итого:			11		
14					12		
15					13		
16					14		
17					15		
18					16		
19					17		
20					18		
21					19		
22					20		
23					21		
24					22		
25					23		
26					24		
27					25		
28					26		
29					27		
30					28		
31					29		
32					30		
33							

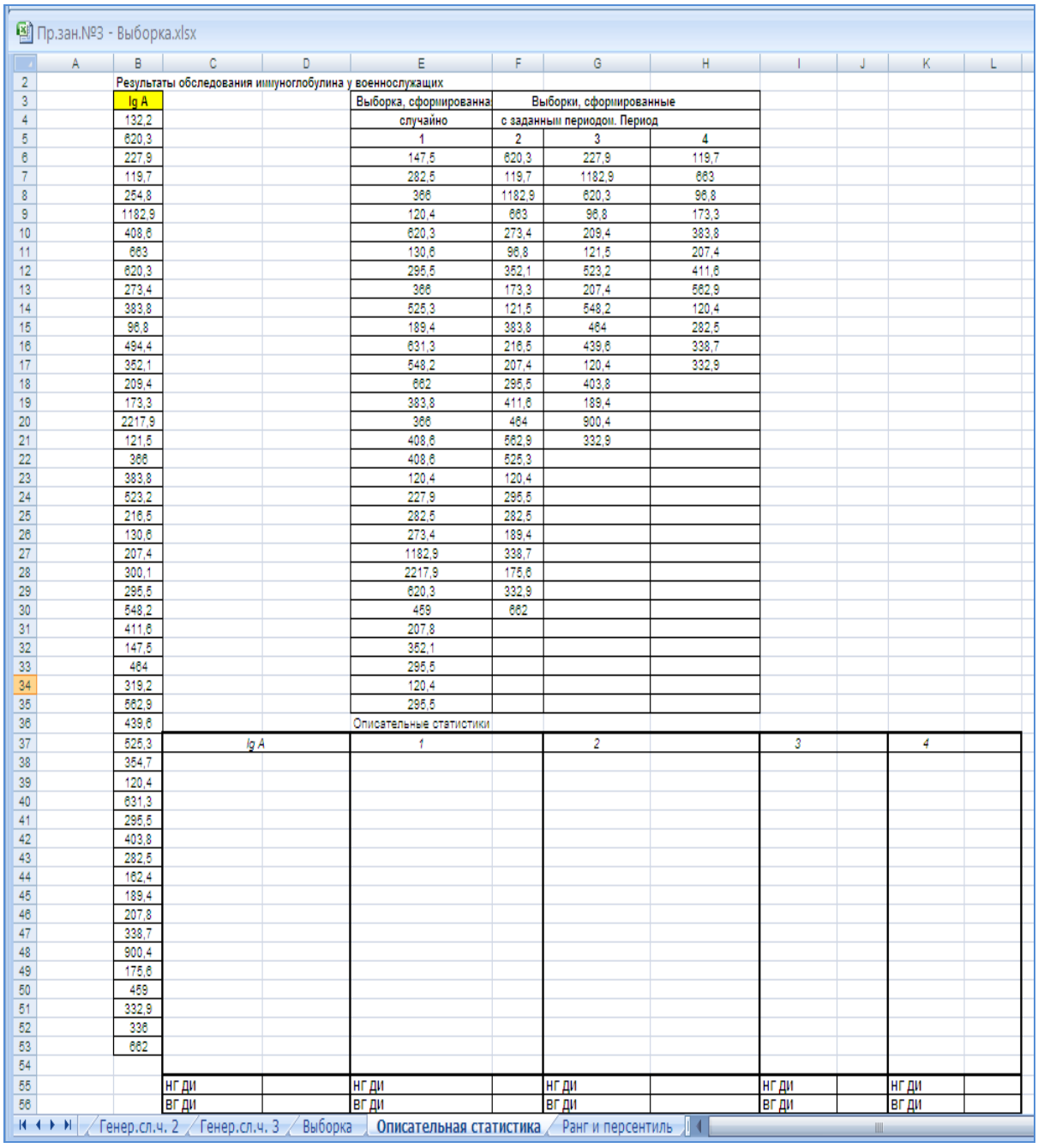
Генер.сл.ч. 1 | Генер.сл.ч. 2 | Генер.сл.ч. 3 | Выборка | Описательная с


Пр.зан.№3 - Выборка.xlsx
карманы
-3
-2
-1
0
1
2
3
4
Среднее
Стандартное отклонение
Медиана
Мода
Генер.сл.ч. 1
Генер.сл.ч. 2
Генер.сл.ч. 3
Выборка
Описательная статис

Пр.зан.№3 - Выборка.xlsx

	A	B	C	D	E	F	G	H
1				ВЫБОРКА				
2		Результаты обследования иммуноглобулина у военнослужащих						
3		Ig A						
4		132,2			Выборка, сформированная	Выборки, сформированны		
5		620,3			случайно	с заданным периодом. Период		
6		227,9			1	2	3	4
7		119,7						
8		254,8						
9		1182,9						
10		408,6						
11		663						
12		620,3						
13		273,4						
14		383,8						
15		96,8						
16		494,4						
17		352,1						
18		209,4						
19		173,3						
20		2217,9						
21		121,5						
22		366						
23		383,8						
24		523,2						
25		216,5						
26		130,6						
27		207,4						
28		300,1						
29		295,5						
30		548,2						
31		411,6						
32		147,5						
33		464						
34		319,2						
35		562,9						
36		439,6						
37		525,3						
38		354,7						
39		120,4						
40		631,3						
41		295,5						
42		403,8						
43		282,5						
44		162,4						
45		189,4						
46		207,8						
47		338,7						
48		900,4						
49		175,6						
50		459						
51		332,9						
52		336						
53		662						
54	Среднее							
55	Медиана							
56	Мода							
57	Стандартное отклонение							
58								

Генер.сл.ч. 1 | Генер.сл.ч. 2 | Генер.сл.ч. 3 | **Выборка** | Описательная статист



Пр.зан.№3 - Выборка.xlsx

Результаты обследования иммуноглобулина у военнослужащих

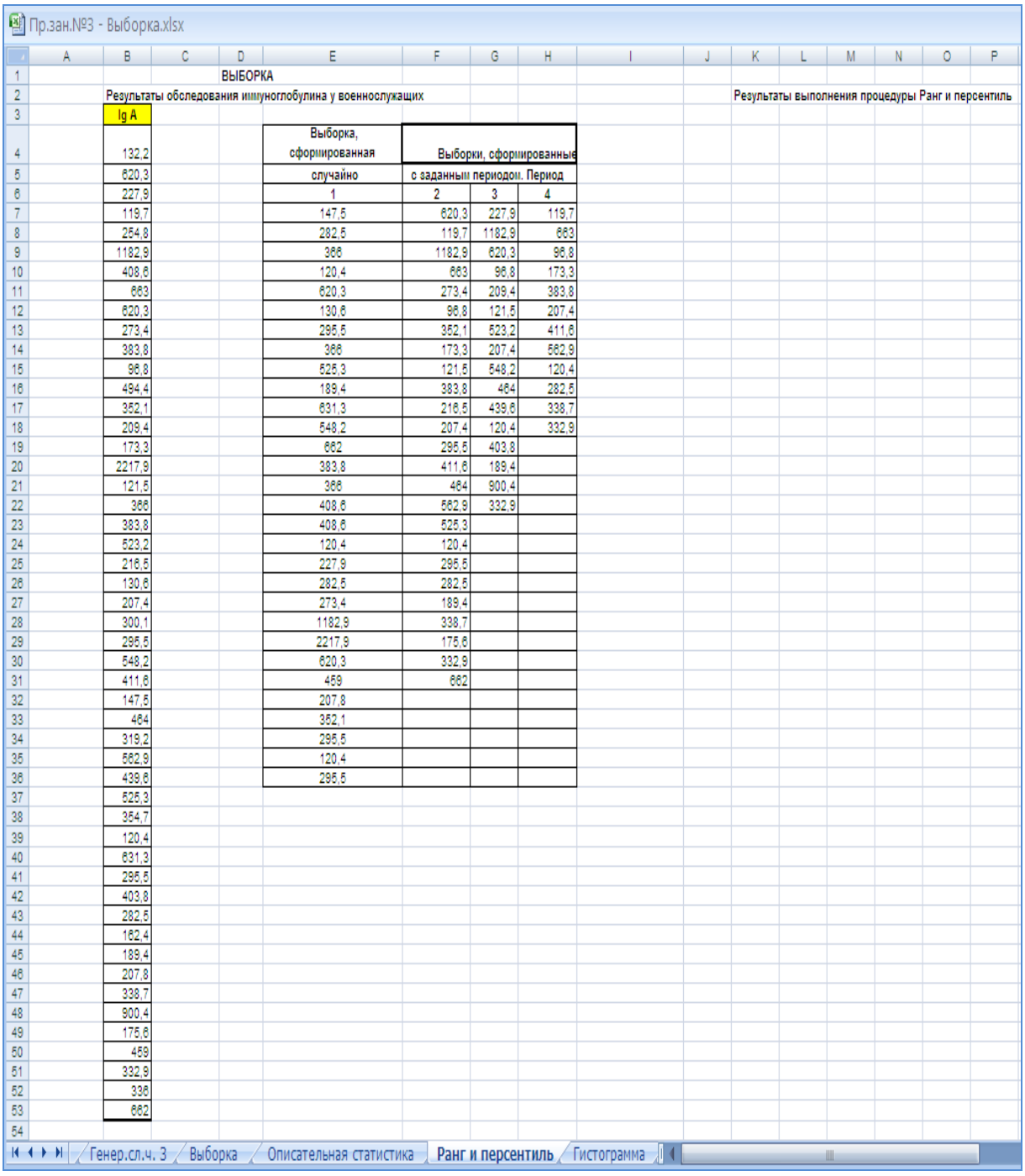
lg A
132,2
620,3
227,9
119,7
254,8
1182,9
408,6
663
620,3
273,4
383,8
96,8
484,4
352,1
209,4
173,3
2217,9
121,5
366
383,8
523,2
216,5
130,6
207,4
300,1
295,5
548,2
411,6
147,5
464
319,2
562,9
439,6
525,3
354,7
120,4
631,3
295,5
403,8
282,5
162,4
189,4
207,8
338,7
900,4
175,6
459
332,9
336
662

Выборка, сформированна	Выборки, сформированные		
случайно	с заданным периодом. Период		
1	2	3	4
147,5	620,3	227,9	119,7
282,5	119,7	1182,9	663
366	1182,9	620,3	96,8
120,4	663	96,8	173,3
620,3	273,4	209,4	383,8
130,6	96,8	121,5	207,4
295,5	352,1	523,2	411,6
366	173,3	207,4	562,9
525,3	121,5	548,2	120,4
189,4	383,8	464	282,5
631,3	216,5	439,6	338,7
548,2	207,4	120,4	332,9
662	295,5	403,8	
383,8	411,6	189,4	
366	464	900,4	
408,6	562,9	332,9	
408,6	525,3		
120,4	120,4		
227,9	295,5		
282,5	282,5		
273,4	189,4		
1182,9	338,7		
2217,9	175,6		
620,3	332,9		
459	662		
207,8			
352,1			
295,5			
120,4			
295,5			

Описательные статистики

lg A		1		2		3		4	
НГ ДИ		НГ ДИ		НГ ДИ		НГ ДИ		НГ ДИ	
ВГ ДИ		ВГ ДИ		ВГ ДИ		ВГ ДИ		ВГ ДИ	

Генер.сл.ч. 2 | Генер.сл.ч. 3 | Выборка | Описательная статистика | Ранг и персентиль



Пр.зан.№3 - Выборка.xlsx

ВЫБОРКА

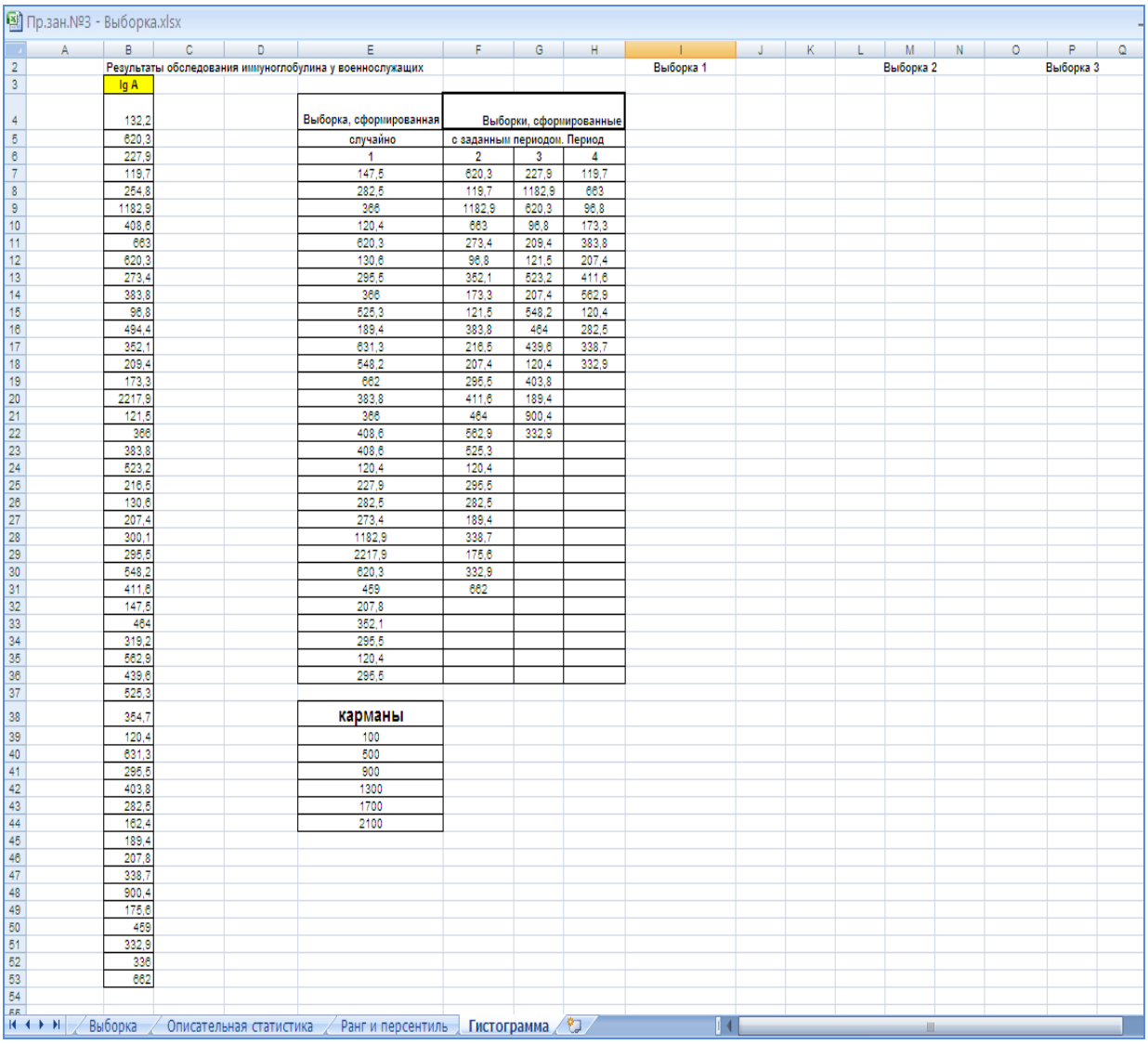
Результаты обследования иммуноглобулина у военнослужащих

Результаты выполнения процедуры Ранг и персентиль

Ig A
132,2
620,3
227,9
119,7
254,8
1182,9
408,6
663
620,3
273,4
383,8
96,8
494,4
352,1
209,4
173,3
2217,9
121,5
366
383,8
523,2
216,5
130,6
207,4
300,1
295,5
548,2
411,6
147,5
464
319,2
562,9
439,6
525,3
354,7
120,4
631,3
295,5
403,8
282,5
162,4
189,4
207,8
338,7
900,4
175,6
459
332,9
336
662

Выборка, сформированная случайно	Выборки, сформированные с заданным периодом. Период		
1	2	3	4
147,5	620,3	227,9	119,7
282,5	119,7	1182,9	663
366	1182,9	620,3	96,8
120,4	663	96,8	173,3
620,3	273,4	209,4	383,8
130,6	96,8	121,5	207,4
295,5	352,1	523,2	411,6
366	173,3	207,4	562,9
525,3	121,5	548,2	120,4
189,4	383,8	464	282,5
631,3	216,5	439,6	338,7
548,2	207,4	120,4	332,9
662	295,5	403,8	
383,8	411,6	189,4	
366	464	900,4	
408,6	562,9	332,9	
408,6	525,3		
120,4	120,4		
227,9	295,5		
282,5	282,5		
273,4	189,4		
1182,9	338,7		
2217,9	175,6		
620,3	332,9		
459	662		
207,8			
352,1			
295,5			
120,4			
295,5			

Генер.сл.ч. 3 | Выборка | Описательная статистика | Ранг и персентиль | Гистограмма



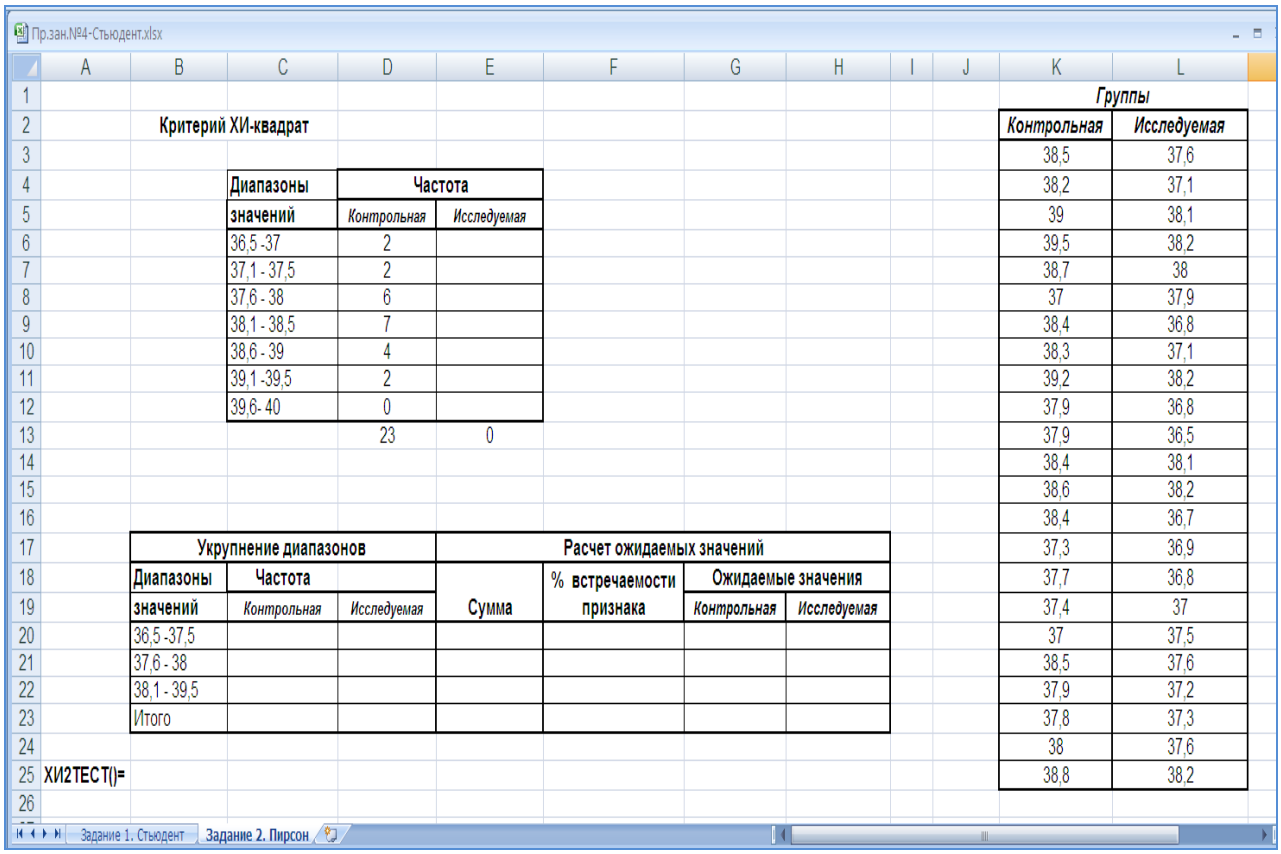
Пр.зан.№3 - Выборка.xlsx

	B	E	F	G	H	I	L	M	O	P
2	Результаты обследования иммуноглобулина у военнослужащих					Выборка 1		Выборка 2		Выборка 3
3	Ig A									
4	132,2	Выборка, сформированная	Выборки, сформированные							
5	620,3	случайно	с заданным периодом. Период							
6	227,9	1	2	3	4					
7	119,7	147,5	620,3	227,9	119,7					
8	254,8	282,5	119,7	1182,9	663					
9	1182,9	366	1182,9	620,3	96,8					
10	408,6	120,4	663	96,8	173,3					
11	663	620,3	273,4	209,4	383,8					
12	620,3	130,6	96,8	121,5	207,4					
13	273,4	295,5	352,1	523,2	411,6					
14	383,8	366	173,3	207,4	662,9					
15	96,8	525,3	121,5	548,2	120,4					
16	494,4	189,4	383,8	464	282,5					
17	352,1	631,3	216,5	439,6	338,7					
18	209,4	548,2	207,4	120,4	332,9					
19	173,3	662	295,5	403,8						
20	2217,9	383,8	411,6	189,4						
21	121,5	366	464	900,4						
22	366	408,6	662,9	332,9						
23	383,8	408,6	525,3							
24	523,2	120,4	120,4							
25	216,5	227,9	295,5							
26	130,6	282,5	282,5							
27	207,4	273,4	189,4							
28	300,1	1182,9	338,7							
29	295,5	2217,9	175,6							
30	548,2	620,3	332,9							
31	411,6	459	662							
32	147,5	207,8								
33	464	352,1								
34	319,2	295,5								
35	662,9	120,4								
36	439,6	295,5								
37	525,3									
38	354,7	карманы								
39	120,4	100								
40	631,3	500								
41	295,5	900								
42	403,8	1300								
43	282,5	1700								
44	162,4	2100								
45	189,4									
46	207,8									
47	338,7									
48	900,4									
49	175,6									
50	459									
51	332,9									
52	336									
53	662									
54										

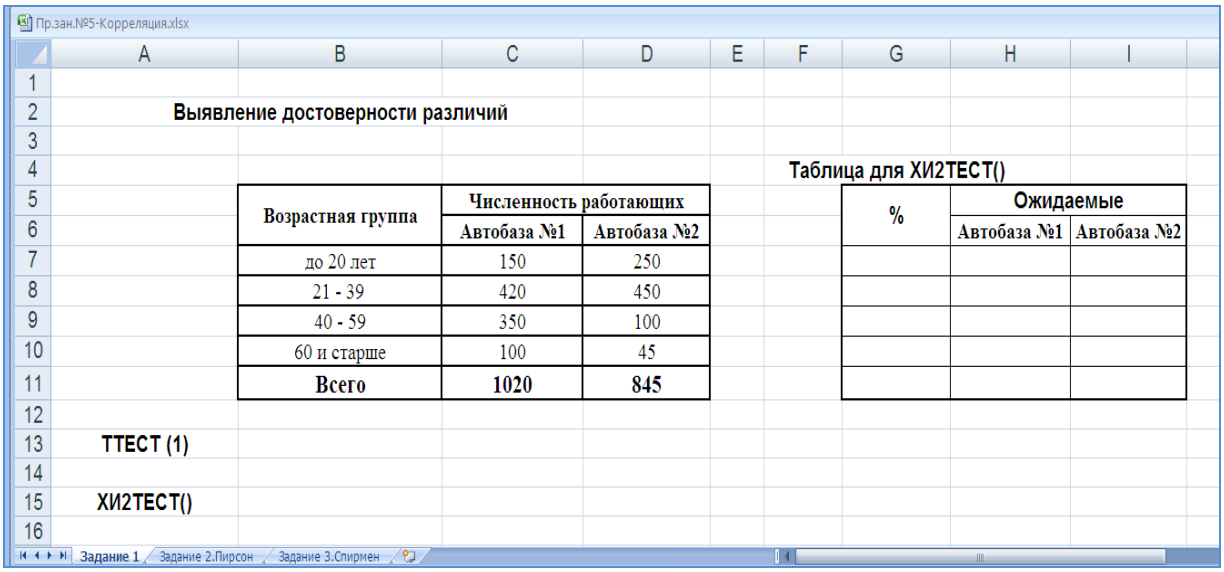
Выборка | Описательная статистика | Ранг и персентиль | Гистограмма

	A	B	C	D	E	F
1		Критерий Стьюдента				
2	Дано:	*Группы*				
3	температура пациентов	*Контрольная*	*Исследуемая*		**Описательные статистики**	
4	в двух группах	38,5	37,6			
5		38,2	37,1			
6		39	38,1			
7		39,5	38,2			
8		38,7	38			
9		37	37,9			
10		38,4	36,8			
11		38,3	37,1			
12		39,2	38,2			
13		37,9	36,8			
14		37,9	36,5			
15		38,4	38,1			
16		38,6	38,2			
17		38,4	36,7			
18		37,3	36,9			
19		37,7	36,8			
20		37,4	37			
21		37	37,5			
22		38,5	37,6			
23		37,9	37,2			
24		37,8	37,3			
25		38	37,6			
26		38,8	38,2			
27	ДИ н.г.для средних					
28	ДИ в.г.для средних					
29						
30	Пациенты одни и те же					
31	ТТЕСТ () Тип=1					
32						
33						
34						
35						
36						
37						
38						
39						
40						
41						
42						
43						
44	Пациенты разные					
45	ТТЕСТ() Тип=3					
46						

Задание 1. Стьюдент | Задание 2. Пирсон


Пр.зан.№4-Стьюдент.xlsx
Критерий ХИ-квадрат
Диапазоны значений
Частота
Контрольная
Исследуемая
36,5 -37
37,1 - 37,5
37,6 - 38
38,1 - 38,5
38,6 - 39
39,1 -39,5
39,6- 40
23
0
Укрупнение диапазонов
Диапазоны значений
Частота
Контрольная
Исследуемая
36,5 -37,5
37,6 - 38
38,1 - 39,5
Итого
Расчет ожидаемых значений
Сумма
% встречаемости признака
Ожидаемые значения
Контрольная
Исследуемая
ХИ2ТЕСТ()=
Группы
Контрольная
Исследуемая
38,5 37,6
38,2 37,1
39 38,1
39,5 38,2
38,7 38
37 37,9
38,4 36,8
38,3 37,1
39,2 38,2
37,9 36,8
37,9 36,5
38,4 38,1
38,6 38,2
38,4 36,7
37,3 36,9
37,7 36,8
37,4 37
37 37,5
38,5 37,6
37,9 37,2
37,8 37,3
38 37,6
38,8 38,2
Задание 1. Стьюдент
Задание 2. Пирсон

ПРИМЕНЕНИЕ ПАРАМЕТРИЧЕСКОГО И НЕПАРАМЕТРИЧЕСКОГО КОРРЕЛЯЦИОННОГО АНАЛИЗА ДАННЫХ



Пр.зан.№5-Корреляция.xlsx

Выявление достоверности различий

Возрастная группа	Численность работающих	
	Автобаза №1	Автобаза №2
до 20 лет	150	250
21 - 39	420	450
40 - 59	350	100
60 и старше	100	45
Всего	**1020**	**845**

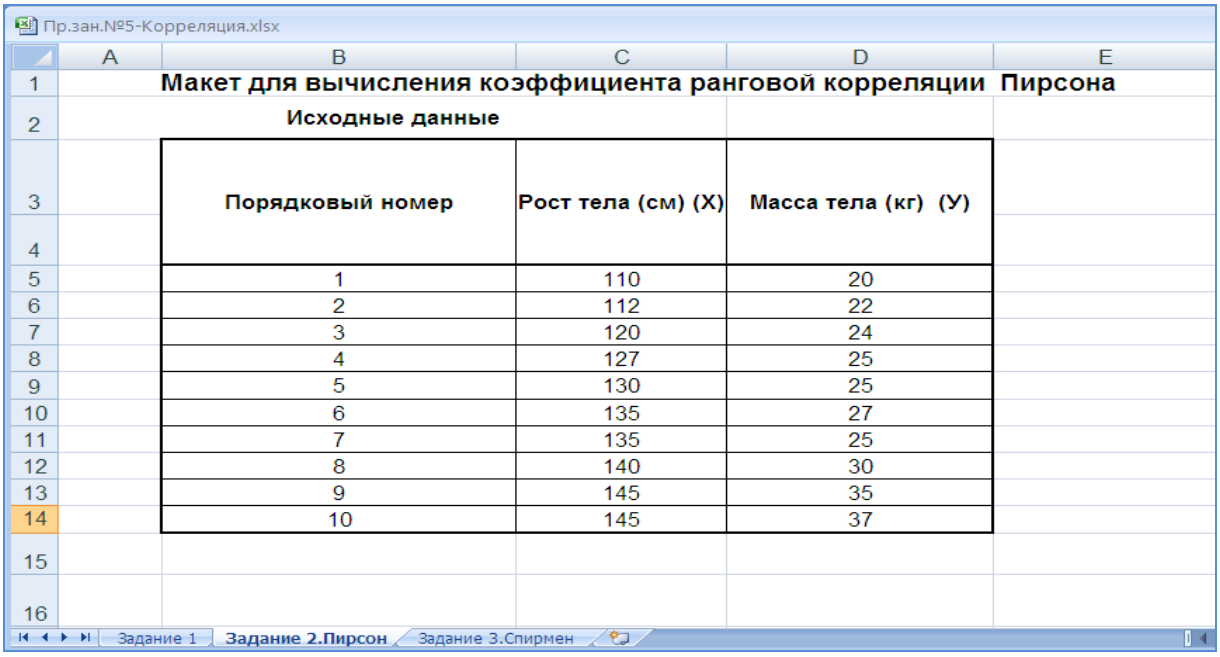
Таблица для ХИ2ТЕСТ()

%	Ожидаемые	
	Автобаза №1	Автобаза №2

ТТЕСТ (1)

ХИ2ТЕСТ()

Задание 1 | Задание 2.Пирсон | Задание 3.Спирмен



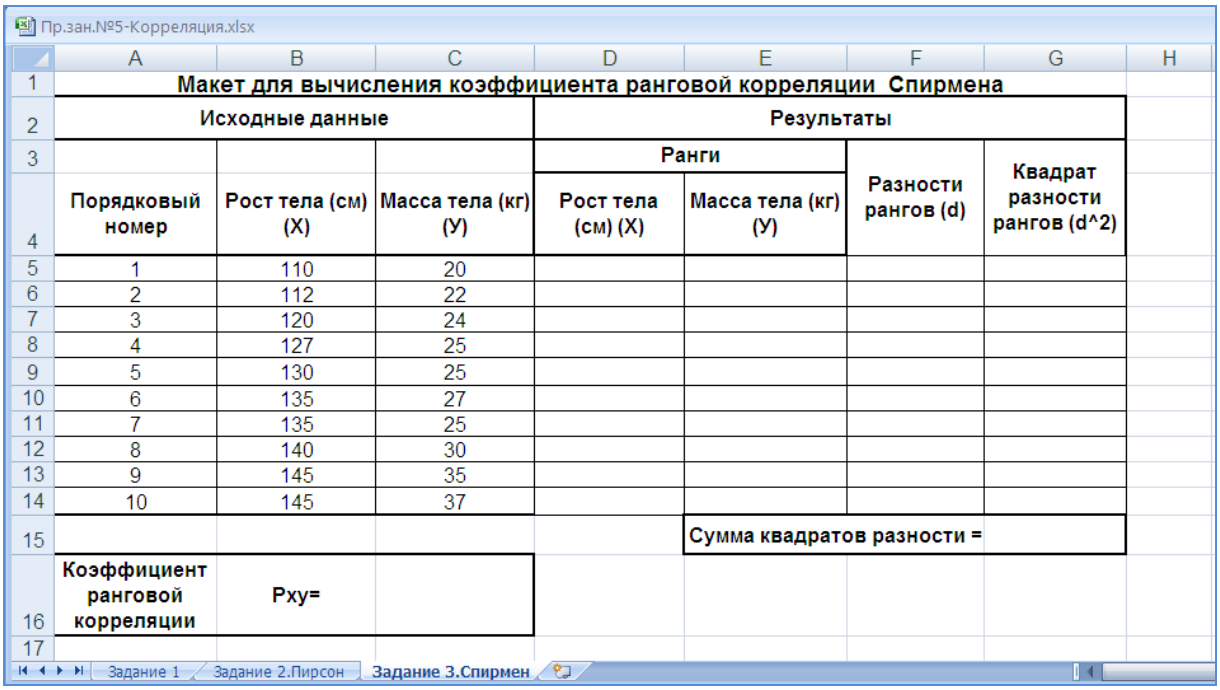
Пр.зан.№5-Корреляция.xlsx

Макет для вычисления коэффициента ранговой корреляции Пирсона

Исходные данные

Порядковый номер	Рост тела (см) (X)	Масса тела (кг) (У)
1	110	20
2	112	22
3	120	24
4	127	25
5	130	25
6	135	27
7	135	25
8	140	30
9	145	35
10	145	37

Задание 1 | Задание 2.Пирсон | Задание 3.Спирмен

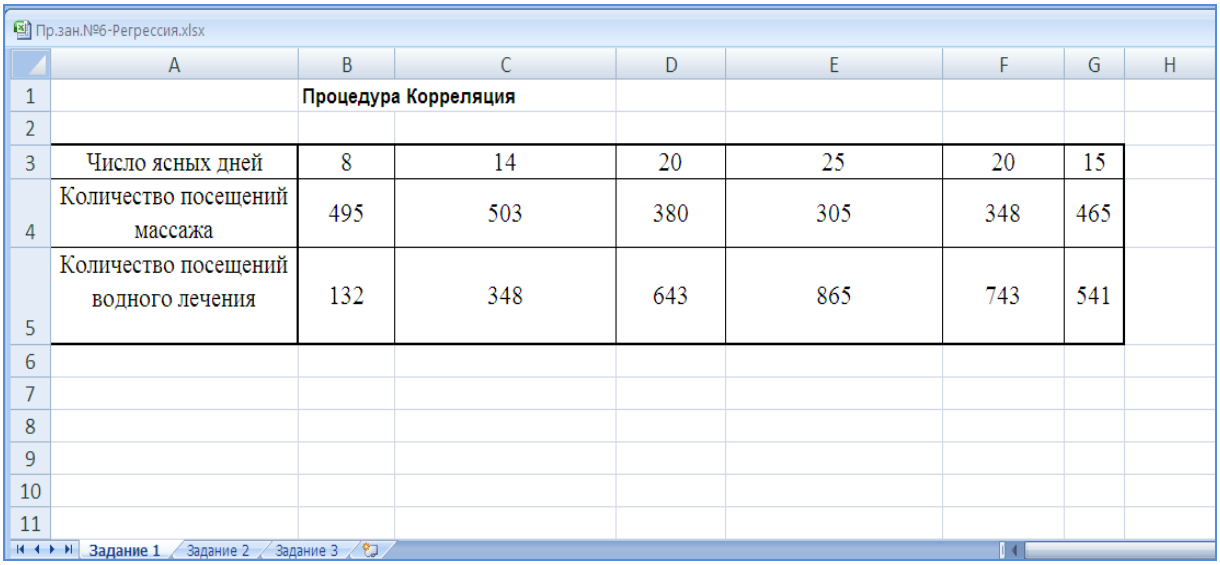


Пр.зан.№5-Корреляция.xlsx

	A	B	C	D	E	F	G	H
1	Макет для вычисления коэффициента ранговой корреляции Спирмена							
2	Исходные данные			Результаты				
3				Ранги		Разности рангов (d)	Квадрат разности рангов (d^2)	
4	Порядковый номер	Рост тела (см) (X)	Масса тела (кг) (У)	Рост тела (см) (X)	Масса тела (кг) (У)			
5	1	110	20					
6	2	112	22					
7	3	120	24					
8	4	127	25					
9	5	130	25					
10	6	135	27					
11	7	135	25					
12	8	140	30					
13	9	145	35					
14	10	145	37					
15					Сумма квадратов разности =			
16	Коэффициент ранговой корреляции	Рху=						
17								

Задание 1 | Задание 2.Пирсон | Задание 3.Спирмен

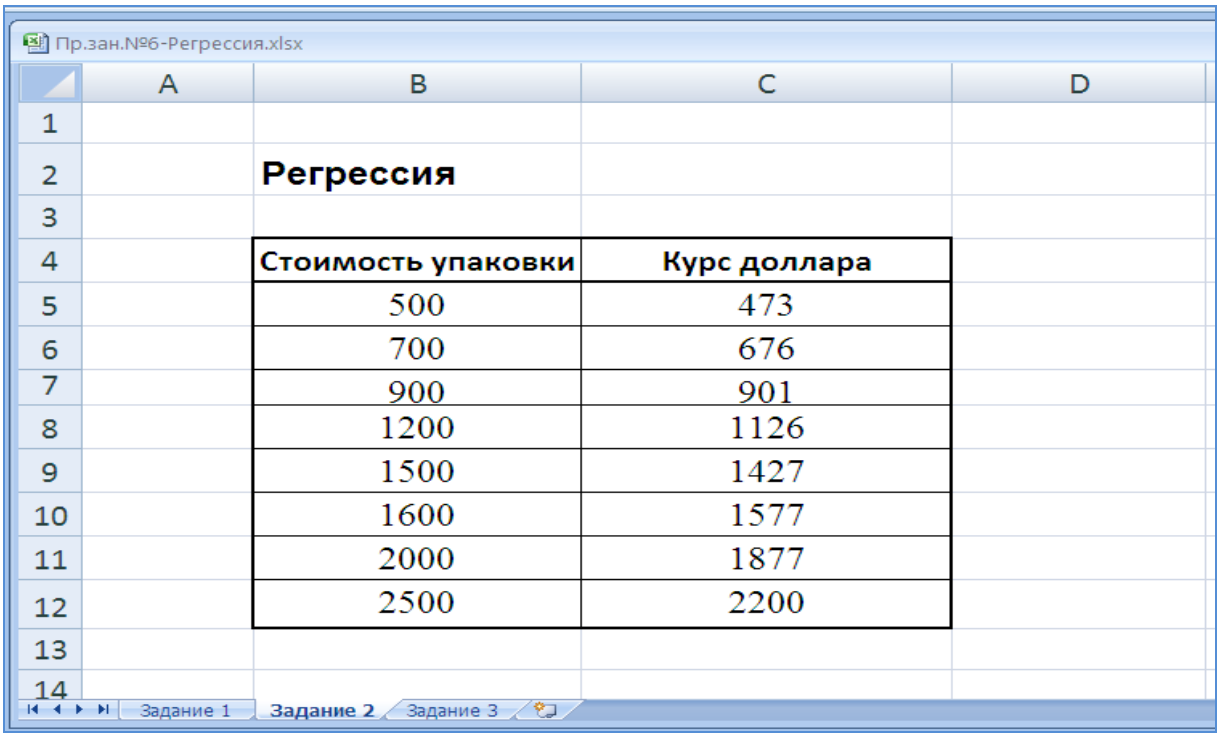
ПРАКТИЧЕСКОЕ ПРИМЕНЕНИЕ РЕГРЕССИОННОГО АНАЛИЗА ДАННЫХ



Пр.зан.№6-Регрессия.xlsx

	A	B	C	D	E	F	G	H
1		Процедура Корреляция						
2								
3	Число ясных дней	8	14	20	25	20	15	
4	Количество посещений массажа	495	503	380	305	348	465	
5	Количество посещений водного лечения	132	348	643	865	743	541	
6								
7								
8								
9								
10								
11								

Задание 1 | Задание 2 | Задание 3



Пр.зан.№6-Регрессия.xlsx

	A	B	C	D
1				
2		**Регрессия**		
3				
4		**Стоимость упаковки**	**Курс доллара**	
5		500	473	
6		700	676	
7		900	901	
8		1200	1126	
9		1500	1427	
10		1600	1577	
11		2000	1877	
12		2500	2200	
13				
14				

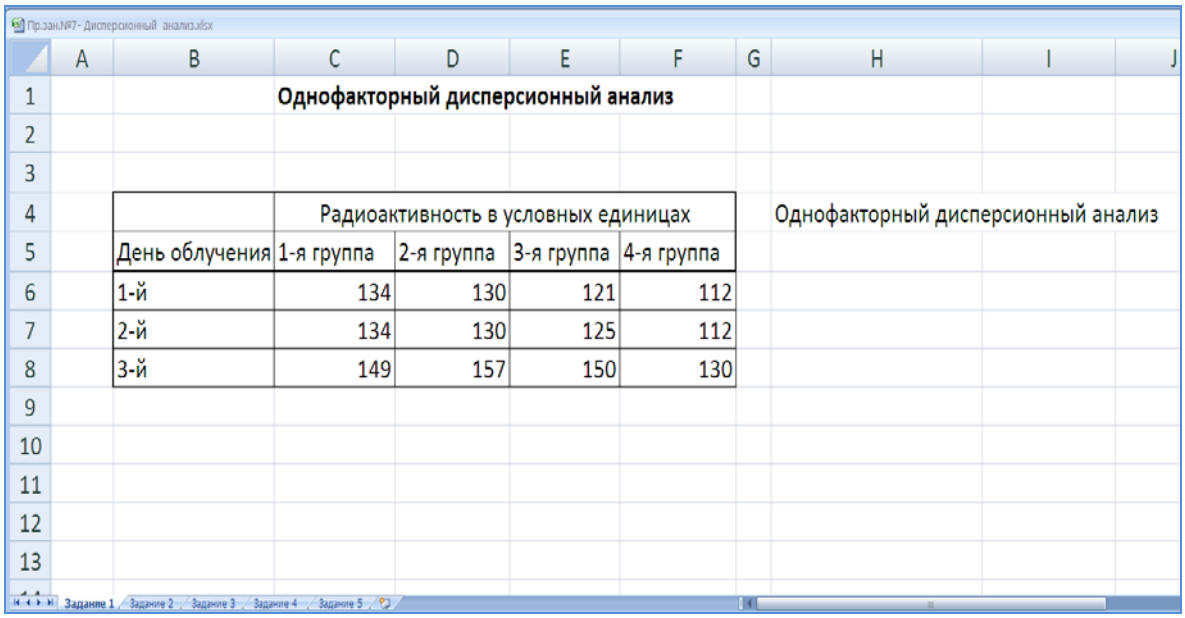
Задание 1 | Задание 2 | Задание 3

Пр.зан.№6-Регрессия.xlsx

	A	B	C	D	E
1		**Регрессия**			
2					
3					
4		X1	X2	У	
5		1	1,3	1160	
6		1	1,3	1155	
7		1,1	1,4	1158	
8		1,1	1,4	1157	
9		1,1	1,5	1160	
10		1,1	1,5	1161	
11		1	1,4	1157	
12		1	1,5	1159	
13		1,2	1,6	1256	
14		1,2	1,7	1260	
15		0,6	1	1040	
16		0,6	1	1039	
17		0,7	1,1	1039	
18		0,7	1,15	1040	
19		0,75	1,2	1040	
20		0,7	1,2	1039	
21		0,7	1,3	1040	
22		0,7	1,3	1039	
23		0,8	1,4	1140	
24		0,8	1,4	1138	
25		0,78	1,5	1240	
26		0,8	1,5	1239	
27		0,78	1,5	1241	
28		0,78	1,6	1240	
29		0,8	1,7	1239	
30		0,8	1,8	1239	
31		0,75	1,8	1240	
32		0,78	1,9	1238	
33		0,75	1,9	1238	
34					

Задание 1 | Задание 2 | **Задание 3**

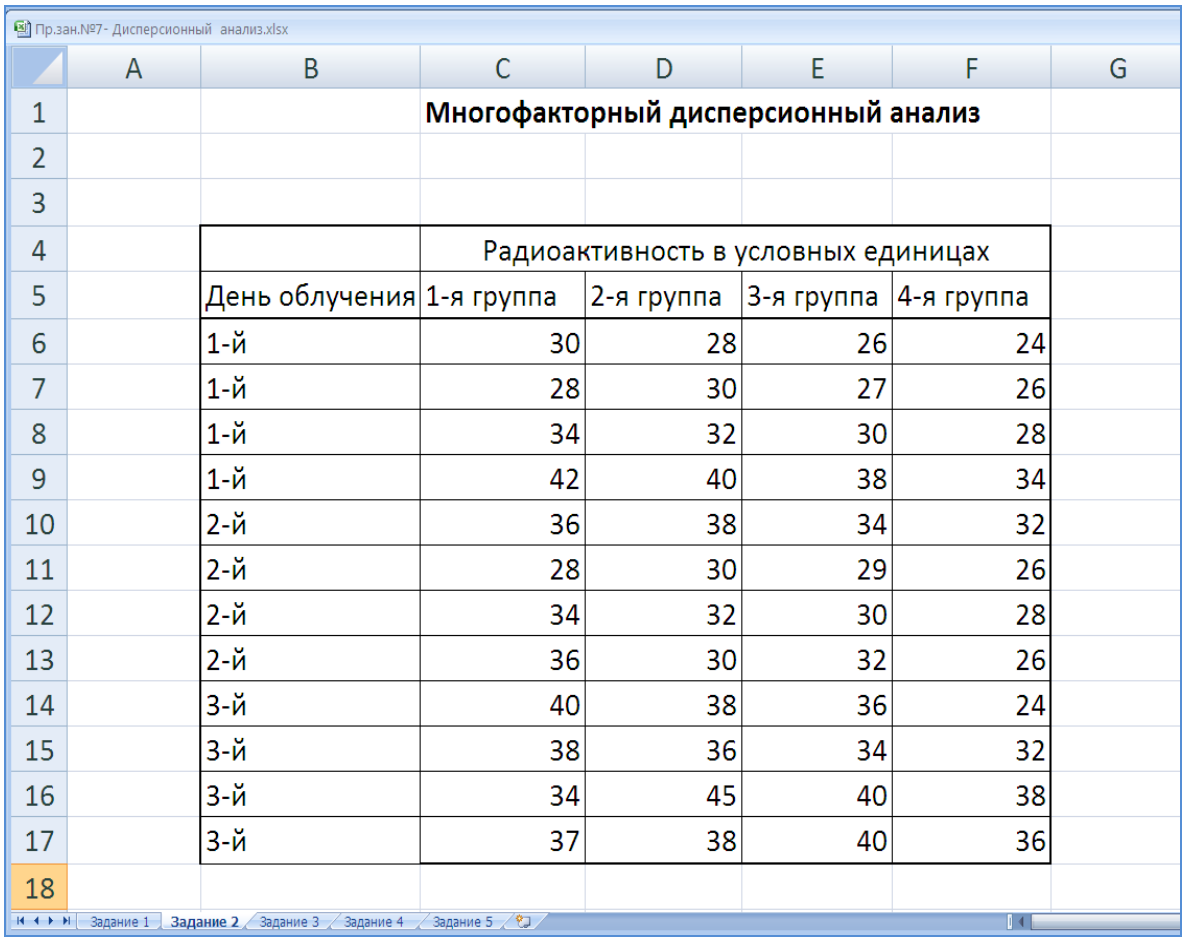
ПРАКТИЧЕСКОЕ ПРИМЕНЕНИЕ ДИСПЕРСИОННОГО АНАЛИЗА ДАННЫХ



Однофакторный дисперсионный анализ

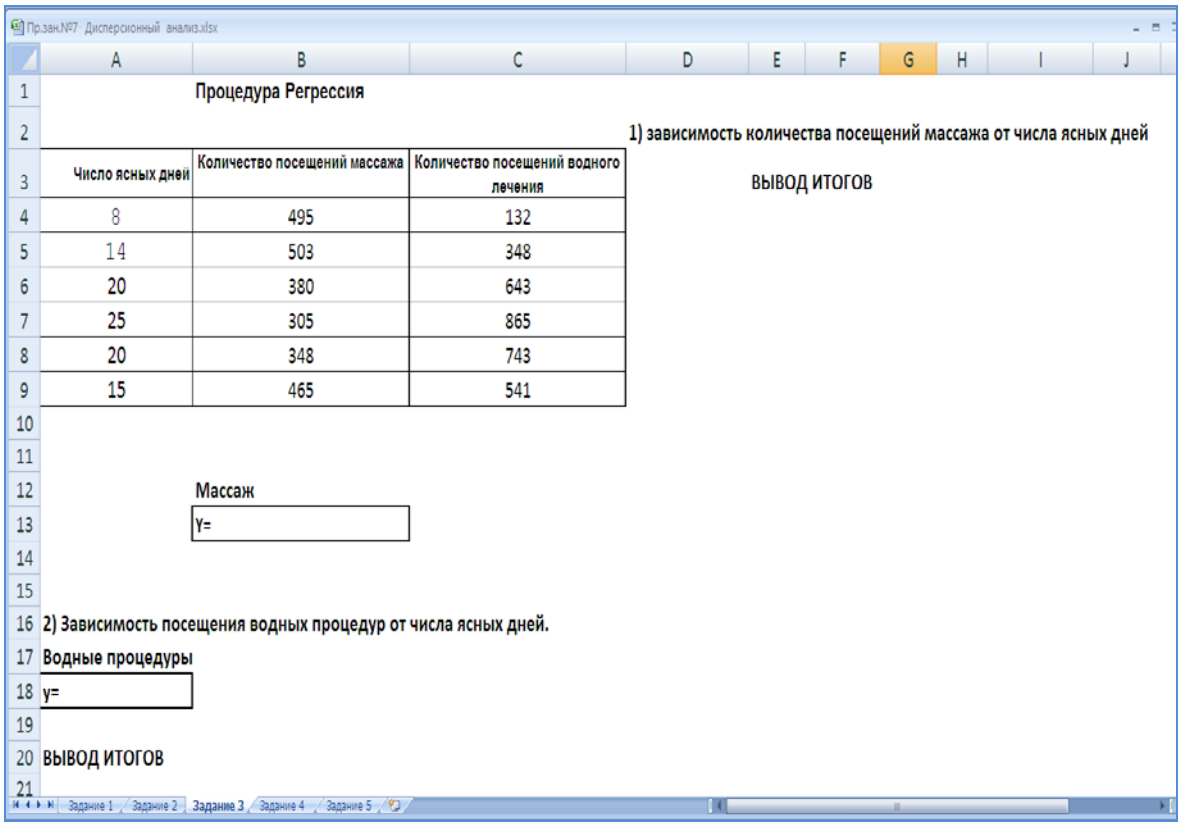
	Радиоактивность в условных единицах			
День облучения	1-я группа	2-я группа	3-я группа	4-я группа
1-й	134	130	121	112
2-й	134	130	125	112
3-й	149	157	150	130

Однофакторный дисперсионный анализ



Многофакторный дисперсионный анализ

	Радиоактивность в условных единицах			
День облучения	1-я группа	2-я группа	3-я группа	4-я группа
1-й	30	28	26	24
1-й	28	30	27	26
1-й	34	32	30	28
1-й	42	40	38	34
2-й	36	38	34	32
2-й	28	30	29	26
2-й	34	32	30	28
2-й	36	30	32	26
3-й	40	38	36	24
3-й	38	36	34	32
3-й	34	45	40	38
3-й	37	38	40	36



Пр.зан.№7 Дисперсионный анализ.xlsx

Процедура Регрессия

1) зависимость количества посещений массажа от числа ясных дней

Число ясных дней	Количество посещений массажа	Количество посещений водного лечения
8	495	132
14	503	348
20	380	643
25	305	865
20	348	743
15	465	541

ВЫВОД ИТОГОВ

Массаж

Y=

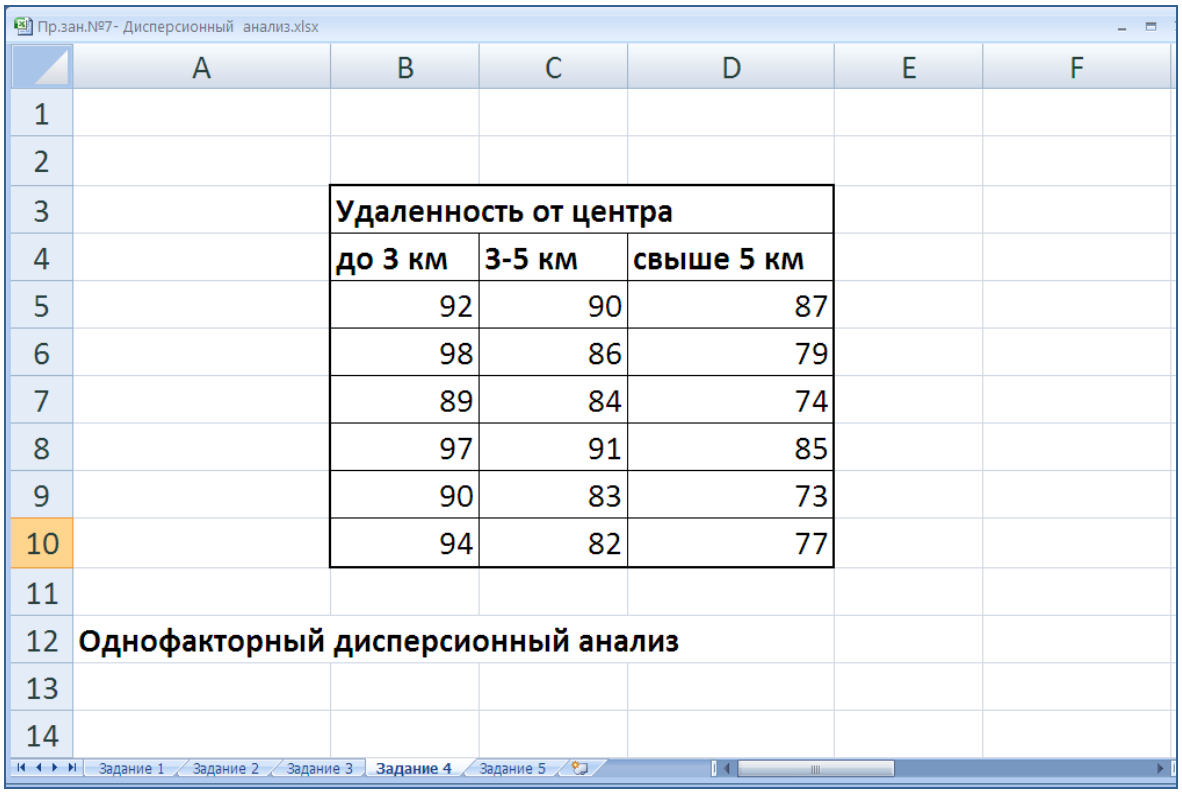
2) Зависимость посещения водных процедур от числа ясных дней.

Водные процедуры

y=

ВЫВОД ИТОГОВ

Задание 1 | Задание 2 | Задание 3 | Задание 4 | Задание 5



Пр.зан.№7- Дисперсионный анализ.xlsx

	Удаленность от центра		
	до 3 км	3-5 км	свыше 5 км
	92	90	87
	98	86	79
	89	84	74
	97	91	85
	90	83	73
	94	82	77

Однофакторный дисперсионный анализ

Задание 1 | Задание 2 | Задание 3 | Задание 4 | Задание 5

Пр.зан.№7- Дисперсионный анализ.xlsx

Двухфакторный дисперсионный анализ с повторениями

Примечание: укажите количество выборок **(число повторений)** 5 - ***это к-во повторений признака ДНИ в выборке***

Среда	Дни	P
лак	10	0,85
лак	10	0,857143
лак	10	0,85
лак	10	0,894737
лак	10	0,882353
лак	15	0,789474
лак	15	0,8
лак	15	0,9375
лак	15	0,833333
лак	15	0,888889
лак	20	0,545455
лак	20	0,625
лак	20	0,521739
лак	20	0,47619
лак	20	0,6
лак	30	0,454545
лак	30	0,444444
лак	30	0,555556
лак	30	0,571429
лак	30	0,5
акр	10	0,888889
акр	10	0,904762
акр	10	0,777778
акр	10	0,736842
акр	10	0,904762
акр	15	0,685714
акр	15	0,741935
акр	15	0,857143
акр	15	0,785714
акр	15	0,727273
акр	20	0,5
акр	20	0,55
акр	20	0,5
акр	20	0,52381
акр	20	0,541667
акр	30	0,45
акр	30	0,380952
акр	30	0,5
акр	30	0,666667
акр	30	0,409091
акфа	10	0,538462
акфа	10	0,428571
акфа	10	0,352941
акфа	10	0,538462
акфа	10	0,181818
акфа	15	0,666667
акфа	15	0,421053
акфа	15	0,285714
акфа	15	0,470588
акфа	15	0,529412
акфа	20	0,423077
акфа	20	0,4
акфа	20	0,454545
акфа	20	0,55
акфа	20	0,416667
акфа	30	0,5
акфа	30	0,36
акфа	30	0,541667

Задание 1 | Задание 2 | Задание 3 | Задание 4 | **Задание 5**

ЛИТЕРАТУРА

Основная:

1. Гараничева, С. Л. Excel для студента-медика [Электронный ресурс] : учеб.-метод. пособие / С. Л. Гараничева. – Витебск : ВГМУ, 2012. – 1 электрон. опт. диск (CD ROM). – Excel для студента-медика. – ISBN 978-985-466-579-5. – 236 с.
2. Гельман, В. Я. Медицинская информатика : практикум / В. Я. Гельман. – 2-е изд. – СПб. : Питер, 2002. – 480 с.
3. Гельман, В. Я. Решение математических задач средствами Excel : практикум / В. Я. Гельман. – СПб. : Питер, 2003. – 237 с.
4. Макарова, Н. В. Статистика в Excel : учеб. пособие / Н. В. Макарова, В. Я. Трофимов. – М. : Финансы и статистика, 2006. – 368 с.
5. Медик, В. А. Математическая статистика в медицине : учеб. пособие / В. А. Медик, М. С. Токмачев. – М. : Финансы и статистика, 2007. – 800 с.
6. Мидлтон, М. Р. Анализ статистических данных с использованием Microsoft Excel для Office XP / М. Р. Мидлтон ; пер. с англ : под. ред. Г. М. Кобелькова. – М. : БИНОМ. Лаборатория знаний, 2005. – 296 с.
7. Омельченко, В. П. Медицинская информатика : учебник / В. П. Омельченко, А. А. Демидова. – М. : ГЭОТАР-Медиа, 2016. – 528 с.
8. Омельченко, В. П. Информатика для врачей: учеб. пособ. / В. П. Омельченко, Н. А. Алексеева. – Ростов н/Д. : Феникс, 2015. – 702 с.
9. Чубарев, В. Н. Фармацевтическая информация / В. Н. Чубарев М. ; под. ред. акад. РАМН А.П. Арзамасцева. – М. : 2000. – 442 с.

Дополнительная:

10. Гараничева, С.Л. Практикум по информатике : учеб. пособие / С. Л. Гараничева. – Витебск: ВГМУ, 2000. – 168 с.
11. Глушанко, В.С. Основы медицинской статистики : учеб.-метод. пособие / В.С. Глушанко [и др.]. – Витебск : ВГМУ, 2012. – 155 с.
12. Жижин, К.С. Медицинская статистика : учебное пособие / К.С. Жижин. – Ростов н/Д: Феникс, 2007. – 160 с. (Высшее образование)

ЗАКЛЮЧЕНИЕ

В настоящее время выполнение научно-исследовательских работ по медико-биологической тематике предполагает обязательное применение методов статистического анализа данных. На статистическом анализе данных базируется понятие доказательной медицины, широко распространенное в нашей стране и за рубежом. В связи с этим полученные Вами знания являются чрезвычайно актуальными для врача.

В процессе изучения материалов, представленных в данном учебном пособии, Вы ознакомились с основными этапами статистической обработки данных, базовыми понятиями статистического анализа научились формировать рандомизированные выборки и описывать их свойства. Кроме того, овладели приемами частотного анализа данных, параметрическими и непараметрическими методами выявления достоверности различий, корреляционного анализа данных. При необходимости Вы сможете самостоятельно представить количественную зависимость признаков в виде математического линейного уравнения (выполнить регрессионный анализ), выявить влияние на ход медико-биологического процесса одного или нескольких факторов (средствами дисперсионного анализа).

Знания, умения и навыки, полученные при изучении основ статистики, помогут Вам грамотно выполнить статистическую обработку данных при подготовке научно-исследовательской студенческой работы, вникнуть в суть вопроса, читая статьи в научных журналах.

Изучив основы статистики и получив практические навыки статистической обработки медико-биологических данных средствами электронных таблиц Microsoft Excel и ее надстройки «Анализ данных», Вы сможете впоследствии самостоятельно освоить приемы работы в среде специализированных программ статистического анализа данных, таких как программа Statistica фирмы Statsoft Inc и другие.

Успехов Вам в углублении и совершенствовании знаний по статистической обработке данных, использовании их в учебном процессе медицинского вуза и в профессиональной деятельности специалиста системы здравоохранения.

Коллектив авторов

Учебное издание

Гараничева Светлана Леонидована,
Таллер Вадим Александрович,
Машеро Екатерина Геннадьевна

Основы статистики

Учебно-методическое пособие

Редактор С.Л. Гараничева
Технический редактор И.А. Борисов
Компьютерная верстка Е.Г. Машеро

Подписано в печать 25.03.2019 г. Формат бумаги 64х84 1/16.
Бумага типографская №2. Гарнитура Times New Roman.
Усл. печ. л. 10,19. Уч.-изд. л. 9,47.
Тираж 1000 экз. Заказ № 336.

Издатель и полиграфическое исполнение
УО « Витебский государственный медицинский университет»
ЛП 023330/453 от 30.12.2013
Пр. Фрунзе,27, 210023, г. Витебск.